Topics in

Undergraduate Mathematics

Ryan Joo

This is (still!) an incomplete draft. For corrections and comments, please send an email to the author at ryanjooruian18@gmail.com, or create a pull request at https://github.com/Ryanjoo18/undergrad-maths.

Draft July 4, 2025.

Preface

I began writing this book during the mid-year break in 2023, which was when I had a bit of free time to read up on undergraduate mathematics, and also started learning LaTeX to write this book.

The objective of this book is to serve as a compilation of essential topics at the *undergraduate* level (as well as serving as my personal notes).

Prerequisites

This book is written such that it is accessible to high school students for self-study. No formal prerequisites are required, although some experience with proofs may be helpful.

Presentation

This book follows the typical style of "Definition", "Theorem", etc. As far as possible, I will try to make clear what I define, and what results I wish to show.

For ease of reference, important terms are ***coloured*** when first defined, and are included in the glossary; less important terms are *italicised* when first defined, and are not included in the glossary.

Note on Problem Solving

Mathematics is about problem solving. Pólya outlined the following problem solving cycle in [Pól45].

1. **Understand the problem**

   Ask yourself the following questions:

   - Do you understand all the words used in stating the problem?
   - Is it possible to satisfy the condition? Is the condition sufficient to determine the unknown? Or is it insufficient? Or redundant? Or contradictory?
   - What are you asked to find or show? Can you restate the problem in your own words?
   - Draw a figure. Introduce suitable notation.
   - Is there enough information to enable you to find a solution?

2. **Devise a plan**

   A partial list of heuristics – good rules of thumb to solve problems – is included:

- Guess and check
- Look for a pattern
- Make an orderly list
- Draw a picture
- Eliminate possibilities
- Solve a simpler problem
- Use symmetry
- Use a model
- Consider special cases
- Work backwards
- Use direct reasoning
- Use a formula
- Solve an equation
- Be ingenious

3. **Execute the plan**

   This step is usually easier than devising the plan. In general, all you need is care and patience, given that you have the necessary skills. Persist with the plan that you have chosen. If it continues not to work discard it and choose another. Don't be misled, this is how mathematics is done, even by professionals.

   - Carrying out your plan of the solution, check each step. Can you see clearly that the step is correct? Can you prove that it is correct?

4. **Check and expand**

   Pólya mentions that much can be gained by taking the time to reflect and look back at what you have done, what worked, and what didn't. Doing this will enable you to predict what strategy to use to solve future problems.

   Look back reviewing and checking your results. Ask yourself the following questions:

   - Can you check the result? Can you check the argument?
   - Can you derive the solution differently? Can you see it at a glance?
   - Can you use the result, or the method, for some other problem?

Building on Pólya's problem solving strategy, Schoenfeld [Sch92] came up with the following framework for problem solving, consisting of four components:

1. **Cognitive resources**: the body of facts and procedures at one's disposal.
2. **Heuristics**: 'rules of thumb' for making progress in difficult situations.
3. **Control**: having to do with the efficiency with which individuals utilise the knowledge at their disposal. Sometimes, this is referred to as metacognition, which can be roughly translated as 'thinking about one's own thinking'.

   (a) These are questions to ask oneself to monitor one's thinking.

   - What (exactly) am I doing? [Describe it precisely.] Be clear what I am doing NOW. Why am I doing it? [Tell how it fits into the solution.]

- Be clear what I am doing in the context of the BIG picture – the solution. Be clear what I am going to do NEXT.

(b) Stop and reassess your options when you

- cannot answer the questions satisfactorily [probably you are on the wrong track]; OR
- are stuck in what you are doing [the track may not be right or it is right but it is at that moment too difficult for you].

(c) Decide if you want to

- carry on with the plan,
- abandon the plan, OR
- put on hold and try another plan.

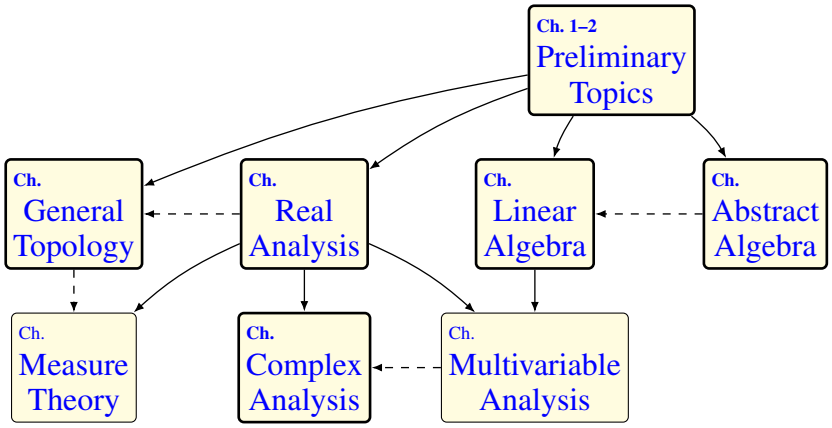
4. **Belief system**: one's perspectives regarding the nature of a discipline and how one goes about working on it.

Acknowledgements

I am greatly thankful to the authors listed in the Bibliography.

Contents

Prerequisite Tree



- A solid arrow ⟶ indicates a required prerequisite, a dotted arrow ⇢ indicates a recommended prerequisite.
- Core topics are in **bold** boxes; other courses (i.e., options or prerequisites) are in light boxes.

Mapping To NUS Courses

The following table is a mapping of Mathematics courses offered at the National University of Singapore (NUS) to the chapters in this book. This is intended for easy reference for NUS students.

Course		Chapters
MA1100T	Basic Discrete Mathematics (T)	1 – 2
MA2001	Linear Algebra I	
MA2002	Calculus	
MA2101/S	Linear Algebra II	3 – 10
MA2104	Multivariable Calculus	
MA2108/S	Mathematical Analysis I	14 – 20
MA2116	Probability	
ST2132	Mathematical Statistics	
MA2202/S	Algebra I	11
MA3210	Mathematical Analysis II	
MA3220	Ordinary Differential Equations	
MA3209	Metric and Topological Spaces	26 –
MA3201	Algebra II	12 – 13
MA3211/S	Complex Analysis I	24 –
MA4211	Functional Analysis	
MA3238	Stochastic Processes	
MA4221	Partial Differential Equations	
MA4262	Measure and Integration	28 –
MA4266	Introduction to Algebraic Topology	
MA5213	Advanced Partial Differential Equations	
MA5206	Graduate Analysis II	
MA5260	Advanced Probability	
MA5205	Graduate Analysis I	

I
Preliminary Topics

1 Mathematical Reasoning and Logic

We will begin with mathematical *language*, the logical connectives and quantifiers, and then we will study the fundamental techniques of *proof*.

1.1 Zeroth-order Logic

Definition 1.1. A ***proposition*** is a sentence which has exactly one truth value, i.e. it is either true or false, but not both and not neither.

A proposition is denoted by uppercase letters such as P and Q. If the proposition P depends on a variable x, it is sometimes helpful to denote it by $P(x)$.

We can do some algebra on propositions:

(i) ***equivalence***, denoted by $P \equiv Q$, means P and Q are logically equivalent statements;

(ii) ***conjunction***, denoted by $P \wedge Q$, means "P and Q";

(iii) ***disjunction***, denoted by $P \vee Q$, means "P or Q";

(iv) ***negation***, denoted by $\neg P$, means "not P".

Here are some useful properties when handling logical statements. You can easily prove all of them using truth tables.

Lemma 1.2.

(i) Double negation law:

$$P \equiv \neg(\neg P)$$

(ii) Commutative laws:

$$P \wedge Q \equiv Q \wedge P$$
$$P \vee Q \equiv Q \vee P$$

(iii) Associative laws:

$$(P \wedge Q) \wedge R \equiv P \wedge (Q \wedge R)$$
$$(P \vee Q) \vee R \equiv P \vee (Q \vee R)$$

(iv) Idempotent laws:

$$P \wedge P \equiv P$$
$$P \vee P \equiv P$$

(v) Distributive laws:

$$P \wedge (Q \vee R) \equiv (P \wedge Q) \vee (P \wedge Q)$$
$$P \vee (Q \wedge R) \equiv (P \vee Q) \wedge (P \vee R)$$

(vi) Absorption laws:

$$P \vee (P \wedge Q) \equiv P$$
$$P \wedge (P \vee Q) \equiv P$$

(vii) de Morgan's laws:

$$\neg(P \vee Q) \equiv (\neg P \wedge \neg Q)$$
$$\neg(P \wedge Q) \equiv (\neg P \vee \neg Q)$$

Remark. Notice that because of the associative laws we can leave out parentheses in statements of the forms $P \wedge Q \wedge R$ and $P \vee Q \vee R$ without ambiguity, because the two possible ways of filling in the parentheses are equivalent.

Statements that are always true are called *tautologies*; for instance $P \vee \neg P$. Similarly, statements that are always false are called *contradictions*; for instance $P \wedge \neg P$.

We can now state a few more useful laws involving tautologies and contradictions.

Lemma 1.3.

(i) Tautology laws: if Q is a tautology, then

$$P \wedge Q \equiv P$$
$$P \vee Q \quad \textit{is a tautology}$$
$$\neg Q \quad \textit{is a contradiction}$$

(ii) Contradiction laws: if Q is a contradiction, then

$$P \vee Q \equiv P$$
$$P \wedge Q \quad \textit{is a contradiction}$$
$$\neg Q \quad \textit{is a tautology}$$

If, only if

We denote an ***implication*** by

$$P \implies Q$$

which means "P implies Q", i.e. if P holds then Q also holds. It is equivalent to saying "If P then Q". $P \implies Q$ is known as a *conditional statement*, where P is known as the *hypothesis* (or *premise*) and Q is known as the *conclusion*.

The only case when $P \implies Q$ is false is when the hypothesis P is true and the conclusion Q is false.

Statements of this form are probably the most common, although they may sometimes appear quite differently. The following all mean the same thing:

(i) if P then Q;

(ii) P implies Q;

(iii) P only if Q;

(iv) P is a sufficient condition for Q;

(v) Q is a necessary condition for P.

Given $P \implies Q$,

- its ***converse*** is $Q \implies P$; both are not logically equivalent;
- its ***inverse*** is $\neg P \implies \neg Q$, i.e. the hypothesis and conclusion of the statement are both negated; both are not logically equivalent;
- the ***contrapositive*** is $\neg Q \implies \neg P$; both are logically equivalent.

To prove $P \implies Q$,

1. assume that P holds,
2. deduce, through some logical steps, that Q holds.

Alternatively, we can prove the contrapositive: assume that Q does not hold, then show that P does not hold.

If and only if, iff

We denote a ***bidirectional implication*** by

$$P \iff Q$$

which means both $P \implies Q$ and $Q \implies P$; $P \iff Q$ is known as a *biconditional statement*. We can read this as "P if and only if Q". The letters "iff" are also commonly used to stand for "if and only if".

$P \iff Q$ is true exactly when P and Q have the same truth value.

These statements are usually best thought of separately as "if" and "only if" statements. To prove $P \iff Q$, prove the statement in both directions:

1. prove $P \implies Q$, and
2. prove $Q \implies P$.

Remember to make very clear, both to yourself and in your written proof, which direction you are doing.

1.2 First-order Logic

The ***universal quantifier*** is denoted by $\forall$, which means "for all" or "for every". A *universal statement* takes the form $\forall x \in X, P(x)$.

The ***existential quantifier*** is denoted by $\exists$, which means "there exists". An *existential statement* takes the form $\exists x \in X, P(x)$, where X is known as the *domain*.

Lemma 1.4 (de Morgan's laws)**.**

$$\neg\left[\forall x \in X, P(x)\right] \equiv \exists x \in X, \neg P(x)$$
$$\neg\left[\exists x \in X, P(x)\right] \equiv \forall x \in X, \neg P(x)$$

To prove a statement of the form $\forall x \in X, P(x)$,

1. Start with "*let $x \in X$ be given*" to address the quantifier with an arbitrary x (this will prove the statement for all $x \in X$).
2. Show that $P(x)$ is true.

Consider statements of the form $\forall x \in X, P(x) \implies Q(x)$; we say that the statement is *vacuously true* if $P(x)$ is false for all $x \in X$.

To prove a statement of the form $\exists x \in X, P(x)$, there is not such a clear steer about how to continue:

- you can construct such an x with the desired properties (constructive proof);
- you can demonstrate logically that such an x must exist because of some earlier assumption (non-constructive proof);
- you can suppose that such an x does not exist, and consequently arrive at some inconsistency (proof by contradiction).

Remark. Read from left to right, and as new elements or statements are introduced they are allowed to depend on previously introduced elements but cannot depend on things that are yet to be mentioned.

Remark. To avoid confusion, it is a good idea to keep to the convention that the quantifiers come first, before any statement to which they relate.

1.3 Methods of Proof

What is a *proof*? Informally, we will define a mathematical proof to be a logical argument that establishes the truth of a mathematical statement. A typical proof proceeds as follows:

1. Start with the given hypotheses.
2. Apply rules of inferences (logical deduction) to get new statements.
3. Repeat Step 2 until we reach the desired conclusion.

We first present some straightforward methods of proof:

- A ***direct proof*** of $P \implies Q$ is a series of valid arguments that start with the hypothesis P and end with the conclusion Q.

$$P \implies \cdots \implies Q$$

- A ***proof by contrapositive*** of $P \implies Q$ is to prove instead $\neg Q \implies \neg P$.
- A ***disproof by counterexample*** is to provide a counterexample to disprove a statement, which makes the negation of the statement true.

  Thus, to disprove $P \implies Q$, the counterexample makes the hypothesis P true, and the conclusion Q false. Likewise, to disprove $\forall x \in X, P(x)$, we prove its negation $\exists x \in X, \neg P(x)$, i.e., find $a \in X$ such that $P(a)$ is false.

  In seeking counterexamples, it is a good idea to keep the cases you consider simple, rather than searching randomly. It is often helpful to consider "extreme" cases; for example, something is zero, a set is empty, or a function is constant.

- A ***proof by cases*** is to first dividing the situation into cases which exhaust all the possibilities, and then show that the statement follows in all cases.

Proof by Contradiction

To ***prove by contradiction***,

1. Assume P is false, i.e., $\neg P$ is true (to prove $P \implies Q$ by contradiction, suppose $P \wedge \neg Q$).
2. Show, through some logical reasoning, that this leads to a contradiction or inconsistency.

   We may arrive at something that contradicts the hypothesis P, or something that contradicts the initial supposition that Q is not true, or we may arrive at something that we know to be universally false.

We illustrate this method of proof using a classic example.

Example (Irrationality of $\sqrt{2}$)**.** Prove that $\sqrt{2}$ is irrational.

Proof. We prove by contradiction. Suppose otherwise, that $\sqrt{2}$ is rational. Then $\sqrt{2} = \frac{a}{b}$ for some $a, b \in \mathbb{Z}, b \neq 0$, a, b coprime.
Squaring both sides gives

$$a^2 = 2b^2.$$

Since RHS is even, LHS must also be even. Hence it follows that a is even. Let $a = 2k$ where $k \in \mathbb{Z}$. Substituting $a = 2k$ into the above equation and simplifying it gives us

$$b^2 = 2k^2.$$

This means that b^2 is even, from which follows again that b is even. This contradicts the assumption that a and b coprime, so we are done. □

Example (Euclid)**.** Prove that there are infinitely many prime numbers.

Proof. Suppose otherwise, that only finitely many prime numbers exist. List them as $p_1, \ldots, p_n$. Consider the number

$$N = p_1 p_2 \cdots p_n + 1.$$

Note that N is divisible by a prime p, yet is coprime to $p_1, \ldots, p_n$. Therefore, p does not belong to our list of all prime numbers, a contradiction. □

Proof of Existence and Uniqueness

To prove existential statements, we can adopt two approaches:

1. ***Constructive proof*** (direct proof)

   To prove statements of the form $\exists x \in X, P(x)$, find or construct *a specific example* for x. To prove statements of the form $\forall y \in Y$, $\exists x \in X, P(x,y)$, construct example for x in terms of y (since x is dependent on y).

   In both cases, you have to justify that your example x

   (a) belongs to the domain X, and

   (b) satisfies the condition P.

2. ***Non-constructive proof*** (indirect proof)

   Use when specific examples are not easy or not possible to find or construct. Make arguments why such objects have to exist. May need to use proof by contradiction. Use definition, axioms or results that involve existential statements.

To ***prove uniqueness*** (after proving existence), we can either

- assume $\exists x, y \in X$ such that $P(x) \wedge P(y)$ is true, then show $x = y$, or
- assume that $\exists x, y \in X$ are distinct such that $P(x) \wedge P(y)$, then derive a contradiction.

We sometimes use $\exists!$ to mean "there exists a unique".

Example. Prove that we can find 100 consecutive positive integers which are all composite numbers.

Proof. We proceed by constructive proof; we will construct integers $n, n+1, n+2, \ldots, n+99$, all of which are composite.

Claim. $n = 101! + 2$.

Then n has a factor of 2 and hence is composite. Similarly, $n+k = 101! + (k+2)$ has a factor $k+2$ and hence is composite for $k = 1, 2, \ldots, 99$.
Hence the existential statement is proven. □

Example. Prove that for all $p, q \in \mathbb{Q}$ with $p < q$, there exists $x \in \mathbb{Q}$ such that $p < x < q$.

Proof. We prove by construction; we want to construct x in terms of p and q, which fulfils the required condition.

Claim. $x = \dfrac{p+q}{2}$.

Evidently $x \in \mathbb{Q}$. Since $p < q$,

$$x = \frac{p+q}{2} < \frac{q+q}{2} = q \implies x < q.$$

Similarly,

$$x = \frac{p+q}{2} > \frac{p+p}{2} = p \implies p < x.$$

Remark. There are two parts to prove: 1) x satisfies the given statement 2) x is within the domain (for this question we do not have to prove x is rational since $\mathbb{Q}$ is closed under addition).

□

Example. Prove that for all rational numbers p and q with $p < q$, there is an irrational number r such that $p < r < q$.

Proof. We prove this by construction. Similarly, our goal is to find an irrational r in terms of p and q.
Note that we cannot simply take $r = \frac{p+q}{2}$; a simple counterexample is the case $p = -1, q = 1$ where $r = 0$ is clearly not irrational.
Since p lies in between p and q, let $r = p + c$ where $0 < c < q - p$. Since $c < q - p$, we have $c = \frac{q-p}{k}$ for some $k > 1$; to make c irrational, we take k to be irrational.

Claim. $r = p + \frac{q-p}{\sqrt{2}}$.

We shall show that (i) $p < r < q$, and (ii) r is irrational.

(i) Since $q - p > 0$, $\frac{q-p}{\sqrt{2}} > 0$ so $r = p + \frac{q-p}{\sqrt{2}} > p + 0 = p$.

$\frac{q-p}{\sqrt{2}} < q - p$ so $r < p + (q - p) = q$.

(ii) We prove by contradiction. Suppose r is rational. We have $\sqrt{2} = \frac{q-p}{r-p}$. Since p, q, r are all rational (and $r - p \neq 0$), RHS is rational. This implies that LHS is rational, i.e. $\sqrt{2}$ is rational, which is a contradiction.

□

Example. Prove that every integer greater than 1 is divisible by a prime.

Proof. We proceed by a non-constructive proof.
If n is prime, then we are done as $n \mid n$.
If n is not prime, then n is composite. So n has a divisor d_1 such that $1 < d_1 < n$. If d_1 is prime then we are done as $d_1 \mid n$. If d_1 is not prime then d_1 is composite, has divisor d_2

such that $1 < d_2 < n$.
If d_2 is prime, then we are done as $d_2 \mid d_1$ and $d_1 \mid n$ imply $d_2 \mid n$. If d_2 is not prime then d_2 is composite, has divisor d_3 such that $1 < d_3 < d_2$.
Continuing in this manner after k times, we will get

$$1 < d_k < d_{k-1} < \cdots < d_2 < d_1 < n$$

where $d_i \mid n$ for all i.
Since there can only be a finite number of d_i's between 1 and n, this process must stop after finite steps. On the other hand, the process will stop only if there is a d_i which is a prime. Hence we conclude that there must be a divisor d_i of n that is prime. □

Remark. This proof is also known as *proof by infinite descent*, a method which relies on the well-ordering principle on $\mathbb{N}$.

Example. Prove that the equation $x^2 + y^2 = 3z^2$ has no solutions (x, y, z) in integers where $z \neq 0$.

Proof. Suppose (x, y, z) is a solution. WLOG assume $z > 0$. By the least integer principle, we may also assume that our solution has z minimal. Taking remainders modulo 3, we see that

$$x^2 + y^2 \equiv 0 \pmod 3$$

Since perfect squares can only be congruent to 0 or 1 modulo 3, we must have $x \equiv y \equiv 0 \pmod 3$. Writing $x = 3a$ and $y = 3b$ for $a, b \in \mathbb{Z}$ gives

$$9a^2 + 9b^2 = 3z^2 \implies 3(a^2 + b^2) = z^2 \implies 3 \mid z^2 \implies 3 \mid z$$

Now let $z = 3c$ and cancel 3's to obtain

$$a^2 + b^2 = 3c^2.$$

We have therefore constructed another solution $(a, b, c) = \left(\frac{x}{3}, \frac{y}{3}, \frac{z}{3}\right)$, but $0 < c < z$ contradicts the minimality of z. □

Proof by Mathematical Induction

Induction is an extremely powerful method of proof used throughout mathematics. It deals with infinite families of statements which come in the form of lists. The idea behind induction is in showing how each statement follows from the previous one on the list – all that remains is to kick off this logical chain reaction from some starting point.

The *well-ordering principle* on $\mathbb{N}$ states the following: every non-empty subset $S \subset \mathbb{N}$ has a smallest element; that is, there exists $m \in S$ such that $m \leq k$ for all $k \in S$.

The *principle of induction* states the following: Let $S \subset \mathbb{N}$. If (i) $1 \in S$, and (ii) $k \in S \implies k+1 \in S$, then $S = \mathbb{N}$.

Lemma 1.5. *The well-ordering principle is equivalent to the principle of induction.*

Proof.

$\boxed{\implies}$ Suppose otherwise, for a contradiction, that S exists with the given properties in the principle of induction, but $S \neq \mathbb{N}$.

Consider the set $\mathbb{N} \setminus S$. Then $\mathbb{N} \setminus S$ is not empty. By the well-ordering principle, $\mathbb{N} \setminus S$ has a least element p. Since $1 \in S$, $1 \notin \mathbb{N} \setminus S$ so $p \neq 1$, thus we must have $p > 1$.

Now consider $p-1$. Since p is the least element of $\mathbb{N} \setminus S$, $p-1 \notin \mathbb{N} \setminus S$ so $p-1 \in S$. But by (ii) of the principle of induction, $p-1 \in S$ implies $p \in S$, which contradicts the fact that $p \in \mathbb{N} \setminus S$.

$\boxed{\impliedby}$ Suppose the principle of induction is true. Then this implies that 1.6 is true, which in turn implies that 1.8 is true. In order to prove the well-ordering of $\mathbb{N}$, we prove the following statement $P(n)$ by strong induction on n: If $S \subset \mathbb{N}$ and $n \in S$, then S has a least element.

The basis step is true, because if $1 \in S$ then 1 is the smallest element of S, since there are no smaller elements of $\mathbb{N}$.

Now suppose that $P(k)$ is true for $k = 1, \ldots, n$. To show that $P(n+1)$ is true, let $S \subset \mathbb{N}$ contain $n+1$. If $n+1$ is the smallest element of S, then we are done. Otherwise, S has a smaller element k, and $P(k)$ is true by the inductive hypothesis, so again S has a smallest element.

Hence by strong induction, $P(n)$ is true for all $n \in \mathbb{N}$. This implies the well-ordering of $\mathbb{N}$, because if S is a non-empty subset of $\mathbb{N}$, then pick $n \in S$. Since $n \in \mathbb{N}$, $P(n)$ is true, and therefore S has a smallest element. □

Theorem 1.6 (Principle of mathematical induction)**.** *Let $P(n)$ be a family of statements indexed by $\mathbb{N}$. Suppose that*

(i) $P(1)$ is true;

(ii) for all $k \in \mathbb{N}$, $P(k) \implies P(k+1)$.

Then $P(n)$ is true for all $n \in \mathbb{N}$.

(i) is known as the *base case*; (ii) is known as the *inductive step*, where we assume $P(k)$ to be true – this is called the *inductive hypothesis* – and show that $P(k+1)$ is true.

Proof. Apply the principle of induction to the set $S = \{n \in \mathbb{N} \mid P(n) \text{ is true}\}$. □

We illustrate the application of this proving technique using a classic example.

Example. Prove that for any $n \in \mathbb{N}$,

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}.$$

Proof. Induct on n. Let $P(n) : \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$.
Clearly $P(1)$ holds. Now suppose $P(k)$ holds for some $k \in \mathbb{N}$, $k \geq 1$; that is,

$$\sum_{i=1}^{k} i = \frac{k(k+1)}{2}.$$

Adding $k+1$ to both sides,

$$\begin{aligned}\sum_{i=1}^{k+1} i &= \frac{k(k+1)}{2} + (k+1) \\ &= \frac{(k+1)(k+2)}{2} \\ &= \frac{(k+1)[(k+1)+1]}{2}\end{aligned}$$

thus $P(k+1)$ is true. Hence by induction, the result holds. □

Example (Bernoulli's inequality)**.** Let $x \in \mathbb{R}$, $x > -1$. Then for all $n \in \mathbb{N}$,

$$(1+x)^n \geq 1 + nx.$$

Proof. Induct on n. Fix $x > -1$. Let $P(n) : (1+x)^n \geq 1+nx$.
The base case $P(1)$ is clear. Suppose that $P(k)$ is true for some $k \in \mathbb{Z}^+$, $k \geq 1$. That is, $(1+x)^k \geq 1+kx$. Note that $1+x > 0$, and $kx^2 \geq 0$ (since $k > 0$ and $x^2 \geq 0$). Then

$$\begin{aligned}(1+x)^{k+1} &= (1+x)(1+x)^k \\ &\geq (1+x)(1+kx) && \text{[induction hypothesis]} \\ &= 1 + (k+1)x + kx^2 \\ &\geq 1 + (k+1)x && [\because kx^2 \geq 0]\end{aligned}$$

so $P(k+1)$ is true. Hence by induction, the result holds. □

A corollary of induction is if the family of statements holds for $n \geq N$, rather than necessarily $n \geq 0$:

Corollary 1.7. *Let $P(n)$ be a family of statements indexed by integers $n \geq N$ for some $N \in \mathbb{Z}$. Suppose that*

(i) $P(N)$ is true;

(ii) for all $k \geq N$, $P(k) \implies P(k+1)$.

Then $P(n)$ is true for all $n \geq N$.

Proof. Apply 1.6 to the statement $Q(n) = P(n+N)$ for $n \in \mathbb{N}$. □

Another variant on induction is when the inductive step relies on some earlier case(s) but not necessarily the immediately previous case.

Theorem 1.8 (Strong induction)**.** *Let $P(n)$ be a family of statements indexed by $\mathbb{N}$. Suppose that*

(i) $P(1)$ is true;

(ii) for all $k \in \mathbb{N}$, $P(1) \wedge \cdots \wedge P(k) \implies P(k+1)$.

Then $P(n)$ is true for all $n \in \mathbb{N}$.

Proof. Let $Q(n)$ be the statement "$P(k)$ holds for $k = 1, \ldots, n$". Then the conditions for the strong form are equivalent to (i) $Q(1)$ holds and (ii) for $n \in \mathbb{N}$, $Q(n) \implies Q(n+1)$. By 1.6, $Q(n)$ holds for all $n \in \mathbb{N}$, and hence $P(n)$ holds for all n. □

Example (Fundamental theorem of arithmetic)**.** Prove that every natural number greater than 1 may be expressed as a product of one or more prime numbers.

Proof. Let $P(n)$ be the statement that n may be expressed as a product of prime numbers.
Clearly $P(2)$ holds, since 2 is itself prime. Let $n \geq 2$ be a natural number and suppose that $P(k)$ holds for all $k < n$.

- If n is prime then it is trivially the product of the single prime number n.
- If n is not prime, then there must exist some $r, s > 1$ such that $n = rs$. By the inductive hypothesis, each of r and s can be written as a product of primes, and therefore $n = rs$ is also a product of primes.

In both cases, $P(n)$ holds. Hence by strong induction, $P(n)$ is true for all $n \in \mathbb{N}$. □

The following is also another variant on induction.

Theorem 1.9 (Cauchy induction). *Let $P(n)$ be a family of statements indexed by $\mathbb{N}_{\geq 2}$. Suppose that*

(i) $P(2)$ *is true;*

(ii) for all $k \geq 2$, $P(k) \implies P(2k)$;

(iii) for all $k \geq 3$, $P(k) \implies P(k-1)$.

Then $P(n)$ is true for all $n \in \mathbb{N}_{\geq 2}$.

Proof. Suppose, for a contradiction, that the desired result is not true.

A smallest integer $k \geq 2$ must exist with $\neg P(k)$, and it is easy to prove that $k \geq 4$.

Then also $\neg P(k+1)$ and one of k and $k+1$ is even and can be written as $2m$ where m is an integer such that $2 \leq m < k$.

Then also $\neg P(m)$ but this contradicts the minimality of k. □

Pigeonhole Principle

Theorem 1.10 (Pigeonhole principle). *If $kn+1$ objects are distributed among n boxes, one of the boxes will contain at least $k+1$ objects.*

Example (IMO 1972). Prove that every set of 10 two-digit integer numbers has two disjoint subsets with the same sum of elements.

Proof. Let S be the set of 10 numbers. It has $2^{10}-2=1022$ subsets that differ from both S and the empty set. They are the "pigeons".
If $A \subset S$, the sum of elements of A cannot exceed $91+92+\cdots+99=855$. The numbers between 1 and 855, which are all possible sums, are the "holes".
Since the number of "pigeons" exceeds the number of "holes", there will be two "pigeons" in the same "hole". Specifically, there will be two subsets with the same sum of elements. Deleting the common elements, we obtain two disjoint sets with the same sum of elements. □

Example (Putnam 2006). Prove that for every set $X=\{x_1,x_2,\ldots,x_n\}$ of n real numbers, there exists a nonempty subset S of X and an integer m such that

$$\left| m+\sum_{x\in S} s \right| \leq \frac{1}{n+1}.$$

Proof. Recall that the fractional part of a real number x is $x-\lfloor x \rfloor$. Consider the fractional parts of the numbers $x_1, x_1+x_2, \ldots, x_1+x_2+\cdots+x_n$.

- If any of them is either in the interval $\left[0, \frac{1}{n+1}\right]$ or $\left[\frac{n}{n+1}, 1\right]$, then we are done.
- If not, consider these n numbers as the "pigeons" and the $n-1$ intervals

$$\left[\frac{1}{n+1}, \frac{2}{n+1}\right], \left[\frac{2}{n+1}, \frac{3}{n+1}\right], \ldots, \left[\frac{n-1}{n+1}, \frac{n}{n+1}\right]$$

  as the "holes". By the pigeonhole principle, two of these sums, say $x_1+x_2+\cdots+x_k$ and $x_1+x_2+\cdots+x_{k+m}$, belong to the same interval. But then their difference $x_{k+1}+\cdots+x_{k+m}$ lies within a distance of $\frac{1}{n+1}$ of an integer, and we are done.

□

Exercises

Exercise 1.1. Negate the statement

$$\text{for all real numbers } x, \text{ if } x > 2, \text{ then } x^2 > 4$$

Solution. In logical notation, this statement is $(\forall x \in \mathbb{R})[x > 2 \implies x^2 > 4]$.

$$\begin{aligned}
&\neg\{(\forall x \in \mathbb{R})[x > 2 \implies x^2 > 4]\} \\
\equiv&(\exists x \in \mathbb{R})\neg[x > 2 \implies x^2 > 4] \\
\equiv&(\exists x \in \mathbb{R})\neg[(x \leq 2) \vee (x^2 > 4)] \\
\equiv&(\exists x \in \mathbb{R})[(x > 2) \wedge (x^2 \leq 4)]
\end{aligned}$$

□

Exercise 1.2. Use the Unique Factorisation Theorem to prove that, if a positive integer n is not a perfect square, then $\sqrt{n}$ is irrational.
[The Unique Factorisation Theorem states that every integer $n > 1$ has a unique standard factored form, i.e. there is exactly one way to express $n = p_1^{k_1} p_2^{k_2} \cdots p_t^{k_t}$ where $p_1 < p_2 < \cdots < p_t$ are distinct primes and $k_1, k_2, \ldots, k_t$ are some positive integers.]

Solution. Prove by contradiction. Suppose n is not a perfect square and $\sqrt{n}$ is rational. Then $\sqrt{n} = \frac{a}{b}$ for some $a, b \in \mathbb{Z}$. Squaring both sides and clearing denominator gives

$$nb^2 = a^2. \tag{$*$}$$

Consider the standard factored forms of n, a and b:

$$n = p_1^{k_1} p_2^{k_2} \cdots p_t^{k_t}$$

$$\begin{aligned}
a = q_1^{e_1} q_2^{e_2} \cdots q_u^{e_u} &\implies a^2 = q_1^{2e_1} q_2^{2e_2} \cdots q_u^{2e_u} \\
b = r_1^{f_1} r_2^{f_2} \cdots r_v^{f_v} &\implies b^2 = r_1^{2f_1} r_2^{2f_2} \cdots r_v^{2f_v}
\end{aligned}$$

i.e. the powers of primes in the standard factored form of a^2 and b^2 are all even integers.

This means the powers k_i of primes p_i in the standard factored form of n are also even by Unique Factorisation Theorem. Note that all p_i appear in the standard factored form of a^2 with even power $2c_i$, because of $(*)$. By UFT, p_i must also appear in the standard factored form of nb^2 with the same even power $2c_i$.

If $p_i \nmid b$, then $k_i = 2c_i$ which is even. If $p_i \mid b$, then p_i will appear in b^2 with even power $2d_i$. So $k_i + 2d_i = 2c_i$, and hence $k_i = 2(c_i - d_i)$, which is again even.

Hence $n = p_1^{k_1} p_2^{k_2} \cdots p_t^{k_t} = \left(p_1^{\frac{k_1}{2}} p_2^{\frac{k_2}{2}} \cdots p_t^{\frac{k_t}{2}} \right)^2$.

Since $\frac{k_i}{2}$ are all integers, $p_1^{\frac{k_1}{2}} p_2^{\frac{k_2}{2}} \cdots p_t^{\frac{k_t}{2}}$ is an integer and n is a perfect square. This contradicts the given hypothesis that n is not a perfect square. □

Exercise 1.3. Prove that for every pair of irrational numbers p and q such that $p < q$, there is an irrational x such that $p < x < q$.

Solution. Consider the average of p and q, i.e., $\frac{p+q}{2}$. Evidently $p < \frac{p+q}{2} < q$.

Since it may not always be the case that $\frac{p+q}{2}$ is irrational (so we cannot immediately take $x = \frac{p+q}{2}$), we need to consider two cases:

Case 1: $\frac{p+q}{2}$ is irrational. Take $x = \frac{p+q}{2}$, and we are done.

Case 2: $\frac{p+q}{2}$ is rational. Let $r = \frac{p+q}{2}$, and consider the average of p and r, i.e., $\frac{p+r}{2}$.

Evidently $p < \frac{p+r}{2} < r < q$. Since p is irrational and r is rational, $\frac{p+r}{2}$ is irrational. In this case, take $x = \frac{3p+q}{4}$.

□

Exercise 1.4. Given n real numbers $a_1, a_2, \ldots, a_n$. Show that there exists an a_i $(1 \leq i \leq n)$ such that a_i is greater than or equal to the mean of the n numbers.

Solution. Prove by contradiction.

Let $\bar{a}$ denote the mean value of the n given numbers. Suppose $a_i < \bar{a}$ for all a_i. Then

$$\bar{a} = \frac{a_1 + a_2 + \cdots + a_n}{n} < \frac{\bar{a} + \bar{a} + \cdots + \bar{a}}{n} = \frac{n\bar{a}}{n} = \bar{a}.$$

We derive $\bar{a} < \bar{a}$, which is a contradiction.

Hence there must be some a_i such that $a_i > \bar{a}$. □

Exercise 1.5. Prove that the following statement is false: there is an irrational number a such that for all irrational number b, ab is rational.

Idea. Prove the negation of the statement: for every irrational number a, there is an irrational number b such that ab is irrational. We shall adopt a constructive proof (note that we can consider multiple cases and construct more than one b).

Solution. Given an irrational number a, let us consider $\dfrac{\sqrt{2}}{a}$. We consider cases:

Case 1: $\dfrac{\sqrt{2}}{a}$ is irrational. Take $b = \dfrac{\sqrt{2}}{a}$. Then $ab = \sqrt{2}$ is irrational.

Case 2: $\dfrac{\sqrt{2}}{a}$ is rational. Then its reciprocal $\dfrac{a}{\sqrt{2}}$ is rational. Since $\sqrt{6}$ is irrational, the product $\left(\dfrac{a}{\sqrt{2}}\right)\sqrt{6} = a\sqrt{3}$ is irrational. Take $b = \sqrt{3}$, which is irrational. Then $ab = a\sqrt{3}$ is irrational.

□

Exercise 1.6. Prove that there are infinitely many prime numbers that are congruent to 3 modulo 4.

Idea. It is not really possible to come up with a direct proof, so we prove by contradiction.

Solution. Suppose, for a contradiction, that there are only finitely many primes that are congruent to 3 modulo 4. Let $p_1, p_2, \ldots, p_m$ be the list of all the primes that are congruent to 3 modulo 4.

Let $M = (p_1 p_2 \cdots p_m)^2 + 2$.

We have the following observation:

(i) $M \equiv 3 \pmod 4$.

(ii) Every p_i divides $M - 2$.

(iii) None of the p_i divides M. [Otherwise, together with (ii), this will imply p_i divides 2, which is impossible.]

(iv) M is not a prime number. [Otherwise, by (i), M is a prime number congruent to 3 modulo 4. But $M \neq p_i$ for all $1 \leq i \leq m$. This contradicts the assumption that $p_1, p_2, \ldots, p_m$ are all the prime numbers congruent to 3 modulo 4.]

From the above discussion, we know that M is a composite number by (iv). So it has a prime factorization $M = q_1 q_2 \cdots q_k$.

Since M is odd, all these prime factors q_j must be odd, and hence q_j must be congruent to either 1 or 3 modulo 4.

By (iii), q_j cannot be any of the p_i. So all q_j must be congruent to 1 modulo 4. Then M, which is the product of q_j, must also be congruent to 1 modulo 4.

This contradicts (i) that M is congruent to 3 modulo 4.

Hence we conclude that there must be infinitely many primes that are congruent to 3 modulo 4. □

Exercise 1.7. Prove that, for any positive integer n, there exists a perfect square m^2 such that $n \leq m^2 \leq 2n$.

Idea. A direct proof by construction is not quite possible, so we prove by contradiction.

Solution. Suppose, for a contradiction, that $n > m^2$ and $(m+1)^2 > 2n$ for some positive integer n, so that there is no square between n and $2n$. Then

$$(m+1)^2 > 2n > 2m^2.$$

Since we are dealing with integers and the inequalities are strict, we get

$$(m+1)^2 \geq 2m^2 + 2$$

which simplifies to

$$0 \geq m^2 - 2m + 1 = (m-1)^2$$

The only value for which this is possible is $m = 1$, but you can eliminate that easily enough. □

Exercise 1.8. Prove that for every positive integer $n \geq 4$,

$$n! > 2^n.$$

Solution. Induct on n. Let $P(n) : n! > 2^n$.

The base case $P(4)$ is clear. Now suppose $P(k)$ is true for some $k \in \mathbb{N}_{\geq 4}$, i.e., $k! > 2^k$. Then

$$(k+1)! = k!(k+1) > 2^k(k+1) > 2^k \cdot 2 = 2^{k+1},$$

so $P(k+1)$ is true. □

Exercise 1.9. Prove by mathematical induction, for $n \geq 2$,

$$\sqrt[n]{n} < 2 - \frac{1}{n}.$$

Solution. Induct on n. Let $P(n) : \sqrt[n]{n} < 2 - \frac{1}{n}$, for $n \geq 2$.

The base case $P(2)$ is clear. Now assume $P(k)$ is true for $k \geq 2, k \in \mathbb{N}$, i.e., $\sqrt[k]{k} < 2 - \frac{1}{k}$, or

$$k < \left(2 - \frac{1}{k}\right)^k.$$

We want to prove that $P(k+1)$ is true; that is,

$$k+1 < \left(2-\frac{1}{k+1}\right)^{k+1}$$

Since $k > 2$, we have

$$\begin{aligned}\left(2-\frac{1}{k+1}\right)^{k+1} &> \left(2-\frac{1}{k}\right)^{k+1} = \left(2-\frac{1}{k}\right)^{k}\left(2-\frac{1}{k}\right)\\ &> k\left(2-\frac{1}{k}\right) = 2k-1 > k-1 \qquad \text{[by inductive hypothesis]}\end{aligned}$$

so $P(k+1)$ is true. □

Exercise 1.10. Prove that, for all integers $n \geq 3$,

$$\left(1+\frac{1}{n}\right)^{n} < n.$$

Solution. For the base case $P(3)$, $\left(1+\frac{1}{3}\right)^{3} = \frac{64}{27} = 2\frac{10}{27} < 3$. Hence $P(3)$ is true.
Assume that $P(k)$ is true for some $k \in \mathbb{N}_{\geq 3}$; that is,

$$\left(1+\frac{1}{k}\right)^{k} < k.$$

Multiplying both sides by $\left(1+\frac{1}{k}\right)$ (to get a $k+1$ in the power),

$$\left(1+\frac{1}{k}\right)^{k}\left(1+\frac{1}{k}\right) = \left(1+\frac{1}{k}\right)^{k+1} < k\left(1+\frac{1}{k}\right) = k+1$$

Since $k < k+1 \iff \frac{1}{k} > \frac{1}{k+1}$,

$$\left(1+\frac{1}{k}\right)^{k+1} > \left(1+\frac{1}{k+1}\right)^{k+1}$$

The rest of the proof follows easily. □

A sequence of integers F_i, where integer $1 \leq i \leq n$, is called the *Fibonacci sequence* if and only if it is defined recursively by $F_1 = 1$, $F_2 = 1$, $F_n = F_{n-1} + F_{n-2}$ for $n > 2$.

Exercise 1.11. Let (a_n) be a sequence of integers defined recursively by the initial conditions $a_1 = 1$, $a_2 = 1$, $a_3 = 3$ and the recurrence relation $a_n = a_{n-1} + a_{n-2} + a_{n-3}$ for $n > 3$.
For all $n \in \mathbb{N}$, prove that

$$a_n \leq 2^{n-1}.$$

Idea. Given the recurrence relation, we may need to use *strong induction*: use $P(k), P(k+1), P(k+2)$ to prove $P(k+3)$, for all $k \in \mathbb{N}$.

Solution. Let $P(n) : a_n \leq 2^{n-1}$.

The base cases $P(1), P(2), P(3)$ are clear. Now assume $P(k), P(k+1), P(k+2)$ are true, for some $k \in \mathbb{N}$. We will show that $P(k+3)$ is true.

By the inductive hypothesis, for $k \in \mathbb{N}$ we have

$$a_k \leq 2^k, \quad a_{k+1} \leq 2^{k+1}, \quad a_{k+2} \leq 2^{k+2}.$$

Then

$$\begin{aligned}
a_{k+3} &= a_k + a_{k+1} + a_{k+2} && \text{[start from recurrence relation]} \\
&\leq 2^k + 2^{k+1} + 2^{k+2} && \text{[use inductive hypothesis]} \\
&= 2^k(1 + 2 + 2^2) \\
&< 2^k(2^3) && \text{[approximation, since } 1 + 2 + 2^2 < 2^3\text{]} \\
&= 2^{k+3}
\end{aligned}$$

which is precisely $P(k+3) : a_{k+3} \leq 2^{k+3}$. □

Exercise 1.12. For $m, n \in \mathbb{N}$, prove that

$$F_{n+m+1} = F_n F_m + F_{n+1} F_{m+1}.$$

Solution. Induct on n. Let $P(n) : F_{n+m+1} = F_n F_m + F_{n+1} F_{m+1}$ for all $m \in \mathbb{N}$ in the cases $k = n$ and $k = n + 1$.

To show that $P(0)$ is true, note that

$$F_{m+1} = F_0 F_m + F_1 F_{m+1}$$

and

$$F_{m+2} = F_1 F_m + F_2 F_{m+1}$$

for all m, as $F_0 = 0$ and $F_1 = F_2 = 1$.

Now assume $P(n)$ is true; that is, for all $m \in \mathbb{N}$,

$$\begin{aligned} F_{n+m+1} &= F_n F_m + F_{n+1} F_{m+1}, \\ F_{n+m+2} &= F_{n+1} F_m + F_{n+2} F_{m+1}. \end{aligned}$$

Then

$$\begin{aligned} F_{n+m+3} &= F_{n+m+2} + F_{n+m+1} \\ &= F_n F_m + F_{n+1} F_{m+1} + F_{n+1} F_m + F_{n+2} F_{m+1} \\ &= (F_n + F_{n+1}) F_m + (F_{n+1} + F_{n+2}) F_{m+1} \\ &= F_{n+2} F_m + F_{n+3} F_{m+1} \end{aligned}$$

thus $P(n+1)$ is true, for all $m \in \mathbb{N}$. □

Exercise 1.13 (NUS MA1100T AY23/24)**.** Suppose $a, b \in \mathbb{N}$. Prove that $\gcd(a,b) = 1$ if and only if $\gcd(ab, a+b) = 1$.

Solution. We first prove a lemma.

Lemma. *Suppose $x, y \in \mathbb{N}$. If there exists $m, n \in \mathbb{Z}$ such that $mx + ny = 1$, then* $\gcd(x,y) = 1$.

Proof. Let $d = \gcd(x,y)$. Then $d \mid x$ and $d \mid y$, so $d \mid mx + ny = 1$. Hence $d \leq 1$. Since $d \geq 1$ by definition, $d = 1$. □

$\boxed{\Longrightarrow}$ Suppose $\gcd(a,b) = 1$. By Bezout's lemma, $ha + kb = 1$ for some $h, k \in \mathbb{Z}$. Then

$$\begin{aligned} 1 &= ha + kb \\ &= (ha + kb)^2 \\ &= h^2a^2 + 2hkab + k^2b^2 \\ &= h^2a^2 + h^2ab + k^2ab + k^2b^2 + 2hkab - h^2ab - k^2ab \\ &= (h^2a + k^2b)(a+b) + (2hk - h^2 - k^2)ab. \end{aligned}$$

By the lemma, $\gcd(ab, a+b) = 1$.

$\boxed{\Longleftarrow}$ Suppose $\gcd(ab, a+b) = 1$. By Bezout's lemma, $p(ab) + q(a+b) = 1$ for some $p, q \in \mathbb{Z}$. Then

$$\begin{aligned} 1 &= pab + q(a+b) \\ &= (pb+q)a + qb. \end{aligned}$$

By the lemma, $\gcd(a,b) = 1$. □

Exercise 1.14 (NUS MA1100 AY24/25)**.** Let $a_1 = 11$, $a_2 = 21$ and $a_{n+1} = 3a_n - 2a_{n-1}$ for all integers n with $n \geq 2$. Prove that for all positive integers n,

$$a_n = 5 \cdot 2^n + 1.$$

Solution. We will prove this by strong induction.

For the base case, we have $a_1 = 5 \cdot 2^1 + 1 = 11$, so it is true. Next, suppose for all $1 \leq k \leq n$, we have $a_k = 5 \cdot 2^k + 1$. We will prove that $a_{k+1} = 5 \cdot 2^{k+1} + 1$. To deduce this, we have

$$\begin{aligned} a_{k+1} &= 3a_k - 2_{k-1} \\ &= 3(5 \cdot 2^k + 1) - 2(5 \cdot 2^{k-1} + 1) \\ &= 15 \cdot 2^k + 3 - 5 \cdot 2^k - 2 \\ &= 10 \cdot 2^k + 1 \\ &= 5 \cdot 2^{k+1} + 1 \end{aligned}$$

By strong induction, the result is true for all positive integers n. □

2 Set Theory

2.1 Basics of Naive Set Theory

Definitions and Notations

A ***set*** S can be loosely defined as a collection of objects[1]. For a set S, we write $x \in S$ to mean that x is an ***element*** of S, and $x \notin S$ if otherwise.

To describe a set, one can list its elements explicitly. A set can also be defined in terms of some property $P(x)$ that the elements $x \in S$ satisfy, denoted by the *set builder notation*:

$$\{x \in S \mid P(x)\}$$

The following sets of numbers are frequently encountered.

- The natural numbers $\mathbb{N} = \{1, 2, 3, \dots\}$ (some people include 0, some do not).
- The integers $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$.
- The rational numbers $\mathbb{Q} = \left\{ \frac{p}{q} \mid p, q \in \mathbb{Z}, q \neq 0 \right\}$.
- The real numbers $\mathbb{R}$ (the construction of which, using Dedekind cuts, will be discussed in Chapter 14).
- The complex numbers $\mathbb{C} = \{x + yi \mid x, y \in \mathbb{R}\}$.

[1] *Russell's paradox*, after the mathematician and philosopher Bertrand Russell (1872–1970), provides a warning as to the looseness of our definition of a set. Suppose H is the collection of sets that are not elements of themselves; that is,

$$H = \{S \mid S \notin S\}.$$

The problem arises when we ask the question of whether or not H is itself in H? On one hand, if $H \notin H$ then H meets the precise criterion for being in H and so $H \in H$, a contradiction. On the other hand, if $H \in H$ then by the property required for this to be the case, $H \notin H$, another contradiction. Thus we have a paradox: H is neither in H, nor not in H.

The modern resolution of Russell's paradox is that we have taken too naive an understanding of "collection", and that Russell's "set" H is in fact not a set. It does not fit within axiomatic set theory (which relies on the so-called ZF axioms), and so the question of whether or not H is in H simply doesn't make sense.

The ***empty set*** $\emptyset$ is the set with no elements.

We say A is a ***subset*** of B if every element of A is in B:

$$A \subset B \quad \text{means} \quad (\forall x)(x \in A \implies x \in B)$$

By construction, we have $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$.

We denote $A \subsetneq B$ to explicitly mean that $A \subset B$ and $A \neq B$; we call A a *proper subset* of B.

Lemma 2.1 ($\subset$ is transitive). *If $A \subset B$ and $B \subset C$, then $A \subset C$.*

Proof. Since $A \subset B$, $x \in A \implies x \in B$. Since $B \subset C$, $x \in B \implies x \in C$. Combining these two implications, we have $x \in A \implies x \in C$. Hence $A \subset C$. □

We say two sets A and B are ***equal*** if and only if they contain the same elements:

$$A = B \quad \text{means} \quad x \in A \iff x \in B.$$

Lemma 2.2 (Double inclusion). *Let $A, B \subset S$. Then*

$$A = B \iff (A \subset B) \wedge (B \subset A)$$

Proof. We have

$$\begin{aligned} A = B &\iff (\forall x)[x \in A \iff x \in B] \\ &\iff (\forall x)[(x \in A \implies x \in B) \wedge (x \in B \implies x \in A)] \\ &\iff \{(\forall x)[x \in A \implies x \in B]\} \wedge (\forall x)[x \in B \implies x \in A)] \\ &\iff (A \subset B) \wedge (B \subset A) \end{aligned}$$

□

Remark. Double inclusion is a useful tool to prove that two sets are equal.

Some frequently occurring subsets of $\mathbb{R}$ are known as ***intervals***, which can be visualised as sections of the real line. We define *bounded intervals*

$$\begin{aligned} (a,b) &= \{x \in \mathbb{R} \mid a < x < b\}, \\ [a,b] &= \{x \in \mathbb{R} \mid a \leq x \leq b\}, \\ [a,b) &= \{x \in \mathbb{R} \mid a \leq x < b\}, \\ (a,b] &= \{x \in \mathbb{R} \mid a < x \leq b\}, \end{aligned}$$

and *unbounded intervals*

$$(a,\infty) = \{x \in \mathbb{R} \mid a < x\},$$
$$[a,\infty) = \{x \in \mathbb{R} \mid a \leq x\},$$
$$(-\infty,a) = \{x \in \mathbb{R} \mid x < a\},$$
$$(-\infty,a] = \{x \in \mathbb{R} \mid x \leq a\}.$$

An interval of the first type (a,b) is called an *open interval*; an interval of the second type $[a,b]$ is called a *closed interval*. Note that if $a = b$, then $[a,b] = \{a\}$, while $(a,b) = [a,b) = (a,b] = \emptyset$.

The ***power set*** $\mathscr{P}(A)$ of A is the set of all subsets of A (including the set itself and the empty set):

$$\mathscr{P}(A) = \{S \mid S \subset A\}.$$

Lemma 2.3. *Let $A,B \subset S$. Then*

$$A \subset B \iff \mathscr{P}(A) \subset \mathscr{P}(B).$$

Proof.

$\boxed{\Longrightarrow}$ Suppose $A \subset B$. Let $X \in \mathscr{P}(A)$. Then $X \subset A$, so $X \subset B$. Thus $X \in \mathscr{P}(B)$. Hence $\mathscr{P}(A) \subset \mathscr{P}(B)$.

$\boxed{\Longleftarrow}$ Suppose A and B are sets such that $\mathscr{P}(A) \subset \mathscr{P}(B)$.

Let $x \in A$. Consider the set $C = \{x\}$. Then $C \subset A$, so $C \in \mathscr{P}(A)$. Thus $C \in \mathscr{P}(B)$, so $C \subset B$.

We know $x \in C$, so we get $x \in B$. Hence $A \subset B$. □

An ***ordered pair*** is denoted by (a,b), where the order of the elements matters. Two pairs (a_1,b_1) and (a_2,b_2) are equal if and only if $a_1 = a_2$ and $b_1 = b_2$. Similarly, we have ordered triples (a,b,c), quadruples (a,b,c,d) and so on. If there are n elements it is called an *n-tuple*.

The ***Cartesian product*** of sets A and B is the set of all ordered pairs with the first element of the pair coming from A and the second from B:

$$A \times B := \{(a,b) \mid a \in A, b \in B\}.$$

More generally, we define

$$\prod_{i=1}^{n} A_i \quad \text{or} \quad A_1 \times \cdots \times A_n$$

to be the set of all ordered n-tuples $(a_1,\ldots,a_n)$, where $a_i \in A_i$ for $1 \leq i \leq n$. If all the A_i are the same, we write the product as A^n.

Later we will define the cartesian product of an *arbitrary* indexed family of sets.

Example. $\mathbb{R}^2$ is the Euclidean plane, $\mathbb{R}^3$ is the Euclidean space, and $\mathbb{R}^n$ is the n-dimensional Euclidean space.

$$\begin{aligned}\mathbb{R}\times\mathbb{R}=\mathbb{R}^2&=\{(x,y)\mid x,y\in\mathbb{R}\}\\ \mathbb{R}\times\mathbb{R}\times\mathbb{R}=\mathbb{R}^3&=\{(x,y,z)\mid x,y,z\in\mathbb{R}\}\\ \mathbb{R}^n&=\{(x_1,x_2,\ldots,x_n)\mid x_1,x_2,\ldots,x_n\in\mathbb{R}\}\end{aligned}$$

Strictly speaking, the Cartesian product is not associative (unless one of the involved sets is empty):

$$(A\times B)\times C\neq A\times(B\times C).$$

For example, if $A=\{1\}$, then $(A\times A)\times A=\{((1,1),1)\}\neq\{(1,(1,1))\}=A\times(A\times A)$.

Lemma 2.4. *Let A,B,C,D be sets.*

(i) $A\times\emptyset=\emptyset\times A=\emptyset$.

(ii) $A\times(B\cup C)=(A\times B)\cup(A\times C)$. *(distributivity)*

(iii) $A\times(B\cap C)=(A\times B)\cap(A\times C)$. *(distributivity)*

(iv) $A\times(B\setminus C)=(A\times B)\setminus(A\times C)$. *(distributivity)*

(v) $(A\cap B)\times(C\cap D)=(A\cap C)\times(B\cap D)$.

(vi) $(A\cup B)\times(C\cup D)\subset(A\cup C)\times(B\cup D)$.

Proof.

(i) Evidently $\emptyset\subset A\times\emptyset$, which is vacuously true.

To show the opposite containment $A\times\emptyset\subset\emptyset$ is equivalent to showing

$$(\forall x)x\in A\times\emptyset\implies x\in\emptyset;$$

but $x\in\emptyset$ is always false, so this is equivalent to

$$(\forall x)\neg(x\in A\times\emptyset).$$

Note that

$$\begin{aligned}x\in A\times\emptyset&\iff(\exists a\in A)(\exists b\in\emptyset)[x=(a,b)]\\&\iff(\exists a)(\exists b)[a\in A\wedge b\in\emptyset\wedge x=(a,b)]\end{aligned}$$

which is always false, since $b\in\emptyset$ is always false.

Hence $(\forall x)\neg(x\in A\times\emptyset)$ holds.

(ii)

$$\begin{aligned} x \in A \times (B \cup C) &\iff x = (a,b) \text{ for some } a \in A, b \in B \cup C \\ &\iff x = (a,b) \text{ for some } a \in A, b \in B \text{ or } b \in C \\ &\iff x = (a,b) \text{ for some } a \in A, b \in B \quad \text{or} \\ &\qquad x = (a,b) \text{ for some } a \in A, b \in C \\ &\iff x \in (A \times B) \text{ or } x \in (A \times C) \\ &\iff x \in (A \times B) \cup (A \times C). \end{aligned}$$

□

Algebra of Sets

We now disuss the algebra of sets. Given $A \subset S$ and $B \subset S$,

(i) The ***union*** $A \cup B$ is the set consisting of elements that are in A or B (or both):

$$A \cup B := \{x \in S \mid x \in A \vee x \in B\}.$$

(ii) The ***intersection*** $A \cap B$ is the set consisting of elements that are in both A and B:

$$A \cap B := \{x \in S \mid x \in A \wedge x \in B\}.$$

We say A and B are ***disjoint*** if both sets have no element in common; that is,

$$A \cap B = \emptyset.$$

More generally, we can take unions and intersections of arbitrary numbers of sets (could be finitely or infinitely many). Given an indexed family of sets $\{A_i\}_{i \in I}$ where I is an *indexing set*, we write

$$\bigcup_{i \in I} A_i = \{x \mid \exists i \in I, x \in A_i\},$$

and

$$\bigcap_{i \in I} A_i = \{x \mid \forall i \in I, x \in A_i\}.$$

When considering families of sets indexed by $\mathbb{N}$, our usual notation will be $\{A_n\}_{n=1}^{\infty}$. In this situation, the notions of *limit superior* and *limit inferior* are sometimes useful:

$$\limsup A_n := \bigcap_{k=1}^{\infty} \bigcup_{n=k} A_n, \quad \liminf A_n := \bigcup_{k=1}^{\infty} \bigcap_{n=k} A_n.$$

You can verify that

$$\limsup A_n = \{x \mid x \in A_n \text{ for infinitely many } n\},$$
$$\liminf A_n = \{x \mid x \in A_n \text{ for all but finitely many } n\}.$$

(iii) The ***complement*** of A, denoted by A^c, is the set containing elements that are not in A:

$$A^c := \{x \in S \mid x \notin A\}.$$

(iv) The ***set difference***, or complement of B in A, denoted by $A \setminus B$, is the subset consisting of those elements that are in A and not in B:

$$A \setminus B := \{x \in A \mid x \notin B\}.$$

Note that $A \setminus B = A \cap B^c$.

Lemma 2.5 (Distributive laws). *Let $A, B, C \subset S$. Then*

(i) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$;

(ii) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

These properties are straightforward consequences of logic laws.

Proof.

(i) We have

$$\begin{aligned} x \in A \cup (B \cap C) &\iff (x \in A) \vee (x \in B \cap C) \\ &\iff (x \in A) \vee [(x \in B) \wedge (x \in C)] \\ &\iff [(x \in A) \vee (x \in B)] \wedge [(x \in A) \vee (x \in C)] \\ &\iff (x \in A \cup B) \wedge (x \in A \cup C) \\ &\iff x \in (A \cup B) \cap (A \cup C). \end{aligned}$$

(ii) Similar to (i).

□

Lemma 2.6 (de Morgan's laws). *Let $A, B \subset S$. Then*

(i) $(A \cup B)^c = A^c \cap B^c$;

(ii) $(A \cap B)^c = A^c \cup B^c$.

Proof.

(i)

$$\begin{aligned} x \in (A \cup B)^c &\iff x \notin A \cup B \\ &\iff x \notin A \quad \wedge \quad x \notin B \\ &\iff x \in A^c \quad \wedge \quad x \in B^c \\ &\iff x \in A^c \cap B^c \end{aligned}$$

(ii) Similar.

□

De Morgan's laws extend naturally to any number of sets. Suppose $\{A_i \mid i \in I\}$ is a family of subsets of S, then

$$\left(\bigcap_{i\in I} A_i\right)^c = \bigcup_{i\in I} {A_i}^c,$$
$$\left(\bigcup_{i\in I} A_i\right)^c = \bigcap_{i\in I} {A_i}^c.$$

Lemma 2.7. *The following hold:*

(i) $\left(\bigcup_{i\in I} A_i\right) \cup B = \bigcup_{i\in I} (A_i \cup B)$

(ii) $\left(\bigcap_{i\in I} A_i\right) \cup B = \bigcap_{i\in I} (A_i \cup B)$

(iii) $\left(\bigcup_{i\in I} A_i\right) \cup \left(\bigcup_{j\in J} B_j\right) = \bigcup_{(i,j)\in I\times J} (A_i \cup B_j)$

(iv) $\left(\bigcap_{i\in I} A_i\right) \cup \left(\bigcap_{j\in J} B_j\right) = \bigcap_{(i,j)\in I\times J} (A_i \cup B_j)$

2.2 Functions

Definitions

Definition 2.8 (Function)**.** A ***function*** $f: X \to Y$ is a mapping of every element of X to some element of Y; we call X and Y the *domain* and *codomain* of f respectively.

Remark. The definition requires that a unique element of the codomain is assigned for every element of the domain. For example, for a function $f: \mathbb{R} \to \mathbb{R}$, the assignment $f(x) = \frac{1}{x}$ is not sufficient as it fails at $x = 0$. Similarly, $f(x) = y$ where $y^2 = x$ fails because $f(x)$ is undefined for $x < 0$, and for $x > 0$ it does not return a unique value; in such cases, we say the the function is *ill-defined*. We are interested in the opposite; functions that are *well-defined*.

If a function is defined on some larger domain than we care about, it may be helpful to restrict the domain:

Definition 2.9 (Restriction)**.** Suppose $f: X \to Y$. The ***restriction*** of f to $A \subset X$ is the map $f|_A : A \to Y$.

Remark. The restriction is almost the same function as the original function – just the domain has changed.

Another rather trivial but nevertheless important function is the identity map:

Definition 2.10 (Identity map)**.** Given a set X, the ***identity*** $\mathrm{id}_X : X \to X$ is defined by

$$\mathrm{id}_X(x) = x \quad (x \in X).$$

Notation. If the domain is unambiguous, the subscript may be omitted.

Injectivity, Surjectivity, Bijectivity

Definition 2.11. Suppose $f: X \to Y$.

(i) f is ***injective*** (or *one-to-one*) if each element of Y has at most one element of X that maps to it:

$$\forall x_1, x_2 \in X, \quad f(x_1) = f(x_2) \implies x_1 = x_2$$

(ii) f is ***surjective*** (or *onto*) if every element of Y is mapped to at least one element of X:

$$\forall y \in Y, \quad \exists x \in X, \quad f(x) = y$$

(iii) f is ***bijective*** if it is both injective and surjective; a bijective function is termed a

bijection.

Notation. We write $X \sim Y$ if there exists a bijection $f : X \to Y$.

Images and Pre-images

Definition 2.12. Suppose $f : X \to Y$. The ***image*** of $A \subset X$ under f is

$$f(A) := \{y \in Y \mid \exists x \in A,\, y = f(x)\}.$$

The ***pre-image*** of $B \subset Y$ under f is

$$f^{-1}(B) := \{x \in X \mid f(x) \in B\}.$$

Notation. In the notation below, unless mentioned otherwise, we will always assume $A \subset X$ and $B \subset Y$.

From the definition, a useful identity is

$$x \in f^{-1}(B) \iff f(x) \in B.$$

Lemma 2.13. *Let $f : X \to Y$.*

(i) $f\left(f^{-1}(B)\right) \subset B$, where equality holds if and only if f is surjective.

(ii) $A \subset f^{-1}(f(A))$, where equality holds if and only if f is injective.

Proof.

(i) Let $y \in f(f^{-1}(B))$. Then $y = f(x)$ for some $x \in f^{-1}(B)$. But $x \in f^{-1}(B)$ is equivalent to $f(x) \in B$. Hence $y \in B$.

(ii) Let $x \in A$. Then $f(x) \in f(A)$ by definition. Hence $x \in f^{-1}(f(A))$.

□

The next result shows that pre-images preserve nice set properties.

Lemma 2.14 (Algebra of pre-images)**.**

(i) If $B_1 \subset B_2$, then $f^{-1}(B_1) \subset f^{-1}(B_2)$. *(preserve inclusions)*

(ii) $f^{-1}\left(\bigcup_{i\in I} B_i\right) = \bigcup_{i\in I} f^{-1}(B_i)$. *(preserve unions)*

(iii) $f^{-1}\left(\bigcap_{i\in I} B_i\right) = \bigcap_{i\in I} f^{-1}(B_i)$. *(preserve intersections)*

(iv) $f^{-1}(B_1 \setminus B_2) = f^{-1}(B_1) \setminus f^{-1}(B_2)$. *(preserve set differences)*

In particular, (iv) implies $f^{-1}(B^c) = \left[f^{-1}(B)\right]^c$.

Proof.

(i) Let $x \in f^{-1}(B_1)$. Then $f(x) \in B_1$, so $f(x) \in B_2$. Hence $x \in f^{-1}(B_2)$.

(ii)

$$\begin{aligned} x \in f^{-1}\left(\bigcup_{i\in I} B_i\right) &\iff f(x) \in \bigcup_{i\in I} B_i \\ &\iff \exists i \in I, \quad f(x) \in B_i \\ &\iff \exists i \in I, \quad x \in f^{-1}(B_i) \\ &\iff x \in \bigcup_{i\in I} f^{-1}(B_i) \end{aligned}$$

(iii)

$$\begin{aligned} x \in f^{-1}\left(\bigcap_{i\in I} B_i\right) &\iff f(x) \in \bigcap_{i\in I} B_i \\ &\iff \forall i \in I, \quad f(x) \in B_i \\ &\iff \forall i \in I, \quad x \in f^{-1}(B_i) \\ &\iff x \in \bigcap_{i\in I} f^{-1}(B_i) \end{aligned}$$

(iv)

$$\begin{aligned} x \in f^{-1}(B_1 \setminus B_2) &\iff f(x) \in B_1 \setminus B_2 \\ &\iff f(x) \in B_1 \quad \text{and} \quad f(x) \notin B_2 \\ &\iff x \in f^{-1}(B_1) \quad \text{and} \quad x \notin f^{-1}(B_2) \\ &\iff x \in f^{-1}(B_1) \setminus f^{-1}(B_2) \end{aligned}$$

□

Unfortunately, images do not behave as nicely as pre-images.

Lemma 2.15 (Algebra of images).

(i) If $A_1 \subset A_2$, *then* $f(A_1) \subset f(A_2)$. *(preserve inclusions)*

(ii) $f\left(\bigcup_{i\in I} A_i\right) = \bigcup_{i\in I} f(A_i)$. *(preserve unions)*

(iii) $f\left(\bigcap_{i\in I} A_i\right) \subset \bigcap_{i\in I} f(A_i)$*, where equality holds if* f *is injective.*

(iv) $f(A_1 \setminus A_2) \supset f(A_1) \setminus f(A_2)$*, where equality holds if* f *is injective.*

In particular, (iv) implies $f(A^c) \supset f(A)^c$, where equality holds if f is surjective.

Proof.

(i) Let $y \in f(A_1)$. Then $y = f(x)$ for some $x \in A_1$. Then $x \in A_2$. By definition, $y \in f(A_2)$.

(ii)

$$\begin{aligned} y \in f\left(\bigcup_{i\in I} A_i\right) &\iff \exists x \in \bigcup_{i\in I} A_i, \quad y = f(x) \\ &\iff \exists i \in I, \quad x \in A_i, \quad y = f(x) \\ &\iff \exists i \in I, \quad y \in f(A_i) \\ &\iff y \in \bigcup_{i\in I} f(A_i). \end{aligned}$$

(iii) Let $x \in \bigcap_{i\in I} A_i$. Then for all $i \in I$, $x \in A_i$, so $f(x) \in f(A_i)$. Thus

$$f(x) \in \bigcap_{i\in I} f(A_i).$$

Hence every $f(x)$ with $x \in \bigcap_{i\in I} A_i$ belongs to $\bigcap_{i\in I} f(A_i)$, which proves the desired result.

(iv) Let $y \in f(A_1) \setminus f(A_2)$. Then $y = f(x)$ for some $x \in A_1$, and $y \notin f(A_2)$ (i.e., for all $z \in A_2$, $f(z) \neq y$).

Suppose, for a contradiction, that $x \in A_2$. Then $f(x) \in f(A_2)$, which contradicts $y \notin f(A_2)$. Thus $x \notin A_2$.

Therefore $x \in A_1 \setminus A_2$, and hence $y = f(x) \in f(A_1 \setminus A_2)$. This yields the desired result.

□

For (iii), equality does not hold in general. Consider the case where f is not injective. **Counterexample:** Let $X = \{1,2\}$, $Y = \{a\}$, and define $f\colon X \to Y$ by $f(1) = f(2) = a$. Let $A_1 = \{1\}$, $A_2 = \{2\}$. Thus

$$f(A_1 \cap A_2) = \emptyset \subset \{a\} = f(A_1) \cap f(A_2)$$

but the inclusion is strict.

For (iv), equality does not hold in general. Consider the case where f is not injective. **Counterexample:** Use the above counterexample for (iii).

Composition

Definition 2.16 (Composition)**.** Given $f\colon X \to Y$ and $g\colon Y \to Z$, the ***composition*** $g \circ f\colon X \to Z$ is defined by

$$(g \circ f)(x) := g(f(x)) \quad (x \in X).$$

The composition of functions is not commutative. However, composition is associative, as the following results shows:

Lemma 2.17. *Suppose $f\colon X \to Y$, $g\colon Y \to Z$, $h\colon Z \to W$. Then*

$$f \circ (g \circ h) = (f \circ g) \circ h.$$

Proof. Let $x \in X$. By definition of composition,

$$(f \circ (g \circ h))(x) = f((g \circ h)(x)) = f(g(h(x))) = (f \circ g)(h(x)) = ((f \circ g) \circ h)(x).$$

□

The next result states that composition preserves injectivity and surjectivity.

Lemma 2.18. *Suppose $f\colon X \to Y$, $g\colon Y \to Z$.*

(i) If f and g are injective, then $g \circ f$ is injective.

(ii) If f and g are surjective, then $g \circ f$ is surjective.

Proof.

(i) Suppose $f\colon X \to Y$ and $g\colon Y \to Z$ are injective.

Suppose $(g \circ f)(x) = (g \circ f)(x')$. Then $g(f(x)) = g(f(x'))$. Injectivity of g implies $f(x) = f(x')$, and injectivity of f implies $x = x'$.

(ii) Suppose $f\colon X \to Y$ and $g\colon Y \to Z$ are surjective.

Let $z \in Z$. By surjectivity of $g\colon Y \to Z$, there exists $y \in Y$ such that $g(y) = z$. By surjectivity of $f\colon X \to Y$, there exists $x \in X$ such that $f(x) = y$.

This means that there exists $x \in X$ such that $g(f(x)) = g(y) = z$, as desired.

□

We now provide a partial converse to the previous result.

Lemma 2.19. *Suppose $f\colon X \to Y$, $g\colon Y \to Z$.*

(i) If $g \circ f$ is injective, then f is injective, but g need not be injective.

(ii) If $g \circ f$ is surjective, then g is surjective, but f need not be surjective.

Proof.

(i) Suppose $f(x) = f(x')$. Then $g(f(x)) = g(f(x')) \implies (g \circ f)(x) = (g \circ f)(x')$. By injectivity of $g \circ f$, this implies $x_1 = x_2$. Hence f is injective.

Let $X = \{1\}$, $Y = \{x, y\}$, $Z = \{z\}$. Define $f\colon X \to Y$ by $f(1) = x$ and $g\colon Y \to Z$ by $g(x) = g(y) = z$.

Then f is (trivially) injective, g is not injective, $g \circ f$ is (vacuously) injective.

(ii) Let $z \in Z$. Since $g \circ f$ is surjective, there exists $x \in X$ such that $g(f(x)) = z$. Let $y = f(x) \in Y$. Then $g(y) = z$. Hence g is surjective.

Let $X = \{1\}$, $B = \{x, y\}$, $C = \{z\}$. Define $f\colon X \to Y$ by $f(1) = x$, $g\colon Y \to Z$ by $g(x) = g(y) = z$. Then f is not surjective, g is surjective, $g \circ f$ is surjective.

□

Proposition 2.20. *$f\colon X \to Y$ is injective if and only if for any set Z and any functions $g_1, g_2\colon Z \to X$,*

$$f \circ g_1 = f \circ g_2 \implies g_1 = g_2.$$

Proof.

$\boxed{\implies}$ Suppose f is injective, and suppose $f \circ g_1 = f \circ g_2$. Let $z \in Z$. Then we have

$$f(g_1(z)) = f(g_2(z)).$$

Injectivity of f implies

$$g_1(z) = g_2(z),$$

so $g_1 = g_2$ (since the choice of $z \in Z$ is arbitrary).

$\boxed{\impliedby}$ Pick $Z = \{1\}$, basically some random one-element set. Then for $x, y \in X$, define

$$g_1\colon Z \to X, \quad g_1(1) = x,$$
$$g_2\colon Z \to Y, \quad g_2(1) = y.$$

Then for $x, y \in X$,

$$f(x) = f(y) \implies f(g_1(1)) = f(g_2(1)) \implies g_1(1) = g_2(1) \implies x = y$$

which shows that f is injective. □

Proposition 2.21. *$f: X \to Y$ is surjective if and only if for any set Z and any functions $g_1, g_2: Y \to Z$,*

$$g_1 \circ f = g_2 \circ f \implies g_1 = g_2.$$

Proof.

$\boxed{\implies}$ Suppose that f is surjective. Let $y \in Y$. Surjectivity of f means there exists $x \in X$ such that $f(x) = y$. Then

$$g_1 \circ f = g_2 \circ f \implies g_1(f(x)) = g_2(f(x)) \implies g_1(y) = g_2(y)$$

so $g_1 = g_2$.

$\boxed{\impliedby}$ We prove the contrapositive. Suppose f is not surjective, then there exists $y \in Y$ such that for all $x \in X$ we have $f(x) \neq y$. We then aim to construct set Z and $g_1, g_2: Y \to Z$ such that

(i) $g_1(y) \neq g_2(y)$

(ii) $\forall y' \neq y, g_1(y') = g_2(y')$

Because if this is satisfied, then $\forall x \in X$, since $f(x) \neq y$ we have from (ii) that $g_1(f(x)) = g_2(f(x))$; thus $g_1 \circ f = g_2 \circ f$, and yet from (i) we have $g_1 \neq g_2$.

We construct $Z = Y \cup \{1,2\}$ for some random $1, 2 \notin Y$.

Then we define

$$g_1: Y \to Z, g_1(y) = 1, g_1(y') = y'$$
$$g_2: Y \to Z, g_2(y) = 2, g_2(y') = y'$$

Then when y is not in the image of f, these two functions will satisfy $g_1 \circ f = g_2 \circ f$ but not $g_1 = g_2$.

So conversely, if for any set Z and any functions $g_i: Y \to Z$ we have $g_1 \circ f = g_2 \circ f \implies g_1 = g_2$, such a value y that is in the codomain but not in the range of f cannot appear, and hence f must be surjective. □

The image and pre-image for composition of functions can be expressed in a very simple way.

Lemma 2.22. *Suppose $f: X \to Y$, $g: Y \to Z$. Let $A \subset X$, $C \subset Z$.*

(i) $(g \circ f)^{-1}(C) = f^{-1}\left(g^{-1}(C)\right)$.

(ii) $(g \circ f)(A) = g(f(A))$.

Proof.

(i)

$$\begin{aligned} x \in (g \circ f)^{-1}(C) &\iff (g \circ f)(x) \in C \\ &\iff g(f(x)) \in C \\ &\iff f(x) \in g^{-1}(C) \\ &\iff x \in f^{-1}\left(g^{-1}(C)\right). \end{aligned}$$

(ii)

$$\begin{aligned} z \in g(f(A)) &\iff \exists y \in f(A), \quad z = g(y) \\ &\iff \exists x \in A, \quad y = f(x), \quad z = g(y) \\ &\iff \exists x \in A, \quad z = g(f(x)) = (g \circ f)(x) \\ &\iff x \in (g \circ f)(A). \end{aligned}$$

□

Invertibility

Recalling that id_X is the identity map on X, we can define invertibility.

Definition 2.23 (Invertibility). Suppose $f\colon X \to Y$. We say that

(i) f is ***left-invertible*** if there exists $g\colon Y \to X$ such that $g \circ f = \mathrm{id}_X$; we call g a *left-inverse* of f;

(ii) f is ***right-invertible*** if there exists $h\colon Y \to X$ such that $f \circ h = \mathrm{id}_Y$; we call h a *right-inverse* of f;

(iii) f is ***invertible*** if there exists $k\colon Y \to X$ which is a left and right inverse of f; we call k an *inverse* of f.

Remark. Notice that if g is left-inverse to f then f is right-inverse to g. A function can have more than one left-inverse, or more than one right-inverse.

Example. Let

$$\begin{aligned} f\colon \mathbb{R} \to [0,\infty), \quad f(x) = x^2 \\ g\colon [0,\infty) \to \mathbb{R}, \quad g(x) = \sqrt{x} \end{aligned}$$

- f is not left-invertible. Suppose otherwise, for a contradiction, that h is a left

inverse of f, so that $hf = \mathrm{id}_{\mathbb{R}}$. Then

Lemma 2.24 (Uniqueness of inverse). *If $f\colon X \to Y$ is invertible, then its inverse is unique.*

Proof. Let g_1 and g_2 be two functions for which $g_i \circ f = \mathrm{id}_X$ and $f \circ g_i = \mathrm{id}_Y$. Using the fact that composition is associative, and the definition of the identity maps, we can write

$$g_1 = g_1 \circ \mathrm{id}_Y = g_1 \circ (f \circ g_2) = (g_1 \circ f) \circ g_2 = \mathrm{id}_X \circ g_2 = g_2.$$

□

Since the inverse is unique, we can give it a notation.

Notation. The inverse of f is denoted by f^{-1}

Remark. Immediately from the definition, if f is invertible then f^{-1} is also invertible, and $(f^{-1})^{-1} = f$.

The following result provides an important and useful criterion for invertibility.

Lemma 2.25 (Invertibility criterion). *Suppose $f\colon X \to Y$. Then*

(i) f is left-invertible if and only if f is injective;

(ii) f is right-invertible if and only if f is surjective;

(iii) f is invertible if and only if f is bijective.

Proof.

(i) $\boxed{\Longrightarrow}$ Suppose f is left-invertible; let g be a left-inverse of f, so $g \circ f = \mathrm{id}_X$.

Now suppose $f(a) = f(b)$. Then applying g to both sides gives $g(f(a)) = g(f(b))$, so $a = b$.

$\boxed{\Longleftarrow}$ Let f be injective. Choose any x_0 in the domain of f. Define $g\colon Y \to X$ as follows; note that each $y \in Y$ is either in the image of f or not.

- If y is in the image of f, it equals $f(x)$ for a *unique* $x \in X$ (uniqueness is because of the injectivity of f), so define $g(y) = x$.
- If y is not in the image of f, define $g(y) = x_0$.

Clearly $g \circ f = \mathrm{id}_X$.

(ii) $\boxed{\Longrightarrow}$ Suppose f is right-invertible; let g be a right-inverse of f, so $f \circ g = \mathrm{id}_Y$.

Let $y \in Y$. Then $f(g(y)) = \mathrm{id}_Y(y) = y$ so $y \in f(X)$. Thus $f(X) = Y$ so f is surjective.

$\boxed{\Longleftarrow}$ Suppose f is surjective. Let $y \in Y$, then y is in the image of f, so we can choose an element $g(y) \in X$ such that $f(g(y)) = y$. This defines a function $g\colon Y \to X$ which is evidently a right-inverse of f.

(iii) $\boxed{\Longrightarrow}$ Suppose f is invertible. Then f is left-invertible and right-invertible. By (i) and (ii), f is injective and surjective, so f is bijective.

$\boxed{\Longleftarrow}$ Suppose f is bijective. Then by (i) and (ii), f has a left-inverse $g\colon Y \to X$ and a right-inverse $h\colon Y \to X$. But "invertible" requires a single function to be *both* a left and right inverse, so we need to show that $g = h$:

$$g = g \circ \mathrm{id}_Y = g \circ (f \circ h) = (g \circ f) \circ h = \mathrm{id}_X \circ h = h$$

so $g = h$ is an inverse of f.

□

The following result shows how to invert the composition of invertible functions.

Proposition 2.26 (Inverse of composition). *Suppose $f\colon X \to Y$, $g\colon Y \to Z$. If f and g are invertible, then $g \circ f$ is invertible, and*

$$(g \circ f)^{-1} = f^{-1} \circ g^{-1}.$$

Proof. Making repeated use of the fact that function composition is associative, and the definition of the inverses f^{-1} and g^{-1}, we note that

$$\begin{aligned}(f^{-1} \circ g^{-1}) \circ (g \circ f) &= ((f^{-1} \circ g^{-1}) \circ g) \circ f \\ &= (f^{-1} \circ (g^{-1} \circ g)) \circ f \\ &= (f^{-1} \circ \mathrm{id}_Y) \circ f \\ &= f^{-1} \circ f \\ &= \mathrm{id}_X\end{aligned}$$

and similarly,

$$\begin{aligned}(g \circ f) \circ (f^{-1} \circ g^{-1}) &= g \circ (f \circ (f^{-1} \circ g^{-1})) \\ &= g \circ ((f \circ f^{-1}) \circ g^{-1}) \\ &= g \circ (\mathrm{id}_Y \circ g^{-1}) \\ &= g \circ g^{-1} \\ &= \mathrm{id}_Z\end{aligned}$$

which shows that $f^{-1} \circ g^{-1}$ satisfies the properties required to be the inverse of $g \circ f$. □

Corollary 2.27. *If $f_1, \ldots, f_n$ are invertible and the composition $f_1 \circ \cdots \circ f_n$ makes sense, then it is also invertible and its inverse is*

$$f_n^{-1} \circ \cdots \circ f_1^{-1}.$$

Proposition 2.28. *$\sim$ is an equivalence relation between sets.*

Proof. We need to prove (i) reflexivity, (ii) symmetry, and (iii) transitivity.

(i) The identity map gives a bijection from a set to itself.

(ii) Suppose $f\colon X \to Y$ is a bijection. Then f is invertible, with inverse $f^{-1}\colon Y \to X$. Since f^{-1} is invertible (with inverse f), it is bijective.

(iii) Suppose $f\colon X \to Y$ and $g\colon Y \to Z$ are bijections, and thus they are invertible. Then by the previous result, $g \circ f$ is invertible and thus bijective.

□

Theorem 2.29 (Cantor–Schröder–Bernstein)**.** *If $f\colon X \to Y$ and $g\colon Y \to X$ are injective, then $A \sim B$.*

2.3 Relations

Definition and Examples

Definition 2.30 (Relation)**.** R is a ***relation*** between A and B if $R \subset A \times B$.
We say $a \in A$ and $b \in B$ are *related* if $(a,b) \in R$, and denote aRb.

Remark. A relation is a set of ordered pairs.

Visually speaking, a relation is uniquely determined by a simple bipartite graph over A and B. On the bipartite graph, this is usually represented by an edge between a and b.

Example. In many cases we do not actually use R to write the relation because there is some other conventional notation:

- The "less than or equal to" relation $\leq$ on the set of real numbers is

$$\{(x,y) \in \mathbb{R}^2 \mid x \leq y\} \subset \mathbb{R}^2;$$

  we write $x \leq y$ if (x,y) is in this set.

- The "divides" relation $\mid$ on $\mathbb{N}$ is

$$\{(m,n) \in \mathbb{N}^2 \mid m \text{ divides } n\} \subset \mathbb{N}^2;$$

  we write $m \mid n$ if (m,n) is in this set.

- For a set S, the "subset" relation $\subset$ on $\mathscr{P}(S)$ is

$$\{(A,B) \in \mathscr{P}(S)^2 \mid A \subset B\} \subset \mathscr{P}(S)^2;$$

  we write $A \subset B$ if (A,B) is in this set.

If $A \times B$ is the smallest Cartesian product of which R is a subset, we call A and B the *domain* and *range* of R respectively, denoted by $\operatorname{dom} R$ and $\operatorname{ran} R$ respectively.

Example. Given $R = \{(1,a),(1,b),(2,b),(3,b)\}$, then $\operatorname{dom} R = \{1,2,3\}$ and $\operatorname{ran} R = \{a,b\}$.

When $A = B$, we have a special name for such relations.

Definition 2.31 (Binary relation)**.** A ***binary relation*** in A is a relation between A and itself; that is, $R \subset A \times A$.

Let A be a set, R a relation on A, $x,y,z \in A$. We say

(i) R is *reflexive* if xRx for all $x \in A$;

(ii) R is *symmetric* if $xRy \implies yRx$;

(iii) R is *anti-symmetric* if xRy and $yRx \implies x = y$;

(iv) R is *transitive* if xRy and $yRz \implies xRz$.

Example (Less than or equal to). The relation $\leq$ on R is reflexive, anti-symmetric, and transitive, but not symmetric.

Definition 2.32. A ***partial order*** on a non-empty set A is a relation on A satisfying reflexivity, anti-symmetry and transitivity.
A ***total order*** on A is a partial order on A such that if for every $x,y \in A$, either xRy or yRx.
A ***well order*** on A is a total order on A such that every non-empty subset of A has a minimal element; that is, for each non-empty $B \subset A$ there exists $s \in B$ such that $s \leq b$ for all $b \in B$.

Equivalence Relations and Partitions

One important type of relation is an equivalence relation. An equivalence relation is a way of saying two objects are, in some particular sense, "the same".

Definition 2.33 (Equivalence relation). A relation $\sim$ on a set A is an ***equivalence relation*** if it is reflexive, symmetric and transitive.

We denote $a \sim b$ for $(a,b) \in R$.

An equivalence relation provides a way of grouping together elements that can be viewed as being the same:

Definition 2.34 (Equivalence class). Given an equivalence relation $\sim$ on a set A, and given $x \in A$, the ***equivalence class*** of x is

$$[x] := \{y \in A \mid y \sim x\}.$$

Grouping the elements of a set into equivalence classes provides a partition of the set, which we define as follows:

Definition 2.35 (Partition). A ***partition*** of a set A is a collection of subsets $\{A_i \mid i \in I\}$ such that

(i) $A_i \neq \emptyset$ for all $i \in I$; (all subsets are non-empty)

(ii) $\bigcup_{i\in I} A_i = A$; (every member of A lies in one of the subsets)

(iii) $A_i \cap A_j = \emptyset$ for every $i \neq j$. (the subsets are disjoint)

Proposition 2.36. *Let $\sim$ be an equivalence relation on a non-empty set X. Then the equivalence classes under $\sim$ are a partition of X.*

To prove this, we need to show that

(i) every equivalence class is non-empty;

(ii) every element of X is an element of an equivalence class;

(iii) every element of X lies in exactly one equivalence class.

Proof.

(i) An equivalence class $[x]$ contains x as $x \sim x$, by reflexivity of the relation. Thus $[x] \neq \emptyset$.

(ii) From (i), note that every $x \in X$ is in the equivalence class $[x]$, so every element of X is an element of at least one equivalence class.

(iii) Suppose, for a contradiction, that some element of X lies in more than one equivalence class. Let $x \in X$ such that $x \in [y]$ and $x \in [z]$; we want to show that $[y] = [z]$ (using double inclusion).

Let $a \in [y]$, so $a \sim y$. Also $x \in [y]$ so $x \sim y$. By symmetry, $y \sim x$. By transitivity, $a \sim x$. Now $x \in [z]$ so $x \sim z$ and similarly $a \sim z$ thus $a \in [z]$. Hence $[y] \subset [z]$.

By the same argument, $[z] \subset [y]$. Hence $[y] = [z]$.

□

Definition 2.37 (Quotient set). The ***quotient set*** is the set of all equivalence classes, denoted by $A/\sim$.

Example (Modular arithmetic). Fix $n \in \mathbb{Z}^+$. Define a relation on $\mathbb{Z}$:

$$a \sim b \iff n \mid (b-a).$$

Lemma. *$\sim$ is a equivalence relation on $\mathbb{Z}$.*

Proof.

(i) Since $n \mid a-a=0$, we have $a \sim a$.

(ii) Suppose $a \sim b$. Then $n \mid a-b$ implies $n \mid b-a$, so $b \sim a$.

(iii) Suppose $a \sim b$ and $b \sim c$. Then $n \mid (a-b)$ and $n \mid (b-c)$, so $n \mid (a-b)+(b-c) = (a-c)$. Thus $a \sim c$.

□

We usually write $a \equiv b \pmod{n}$ if $a \sim b$.
We denote the equivalence class of a by $[a]$, called the *congruence class* of a mod n, which consists of the integers which differ from a by an integral multiple of n; that is,

$$[a] = \{a+kn \mid k \in \mathbb{Z}\}.$$

There are precisely n distinct congruence classes mod n, namely

$$[0], [1], \ldots, [n-1]$$

determined by the possible remainders after division by n; and these congruence classes partition the integers $\mathbb{Z}$. The set of congruence classes is denoted by $\mathbb{Z}/n\mathbb{Z} := \mathbb{Z}/\sim$, called the *integers modulo n*.
Define addition and multiplication on $\mathbb{Z}/n\mathbb{Z}$ as follows: for $[a], [b] \in \mathbb{Z}/n\mathbb{Z}$,

$$[a]+[b] = [a+b]$$
$$[a][b] = [ab].$$

Lemma. *Addition and mulltiplication on $\mathbb{Z}/n\mathbb{Z}$ are well-defined.*

Proof. Suppose $[a_1] = [b_1]$ and $[a_2] = [b_2]$.
Then $a_1 \equiv b_1 \pmod{n}$, or $n \mid (a_1 - b_1)$; let $a_1 = b_1 + sn$ for some integer s. Similarly, let $a_2 = b_2 + tn$ for some integer t.
Then $a_1 + a_2 = (b_1+b_2)+(s+t)n$, so $a_1 + a_2 \equiv b_1 + b_2 \pmod{n}$. Hence $[a_1+a_2] = [b_1+b_2]$.
Similarly, $a_1a_2 = (b_1+sn)(b_2+tn) = b_1b_2 + (b_1t + b_2s + stn)n$ shows that $a_1a_2 \equiv b_1b_2 \pmod{n}$. Hence $[a_1a_2] = [b_1b_2]$. □

Hence we have shown that if $a_1 \equiv b_1 \pmod{n}$ and $a_2 \equiv b_2 \pmod{n}$, then

$$a_1 + a_2 \equiv b_1 + b_2 \pmod{n}, \quad a_1a_2 \equiv b_1b_2 \pmod{n}.$$

An important subset of $\mathbb{Z}/n\mathbb{Z}$ consists of the collection of congruence classes which have a multiplicative inverse in $\mathbb{Z}/n\mathbb{Z}$:

$$(\mathbb{Z}/n\mathbb{Z})^\times := \{[a] \in \mathbb{Z}/n\mathbb{Z} \mid \exists [c] \in \mathbb{Z}/n\mathbb{Z}, [a][c] = [1]\}.$$

Lemma. $(\mathbb{Z}/n\mathbb{Z})^\times$ *equals the collection of congruence classes whose representatives are relatively prime to n:*

$$(\mathbb{Z}/n\mathbb{Z})^\times = \{[a] \in \mathbb{Z}/n\mathbb{Z} \mid (a,n) = 1\}.$$

Example (Rationals)**.** Define a relation on $\mathbb{Z} \times \mathbb{Z}^*$, where $\mathbb{Z}^* = \mathbb{Z} \setminus \{0\}$:

$$(a,b) \sim (c,d) \iff ad = bc.$$

Denote the congruence class of (a,b) by a/b.
Let $\mathbb{Q} := \mathbb{Z} \times \mathbb{Z}^*$, with addition and multiplication defined by

$$\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd}$$
$$\frac{a}{b}\frac{c}{d} = \frac{ac}{bd}$$

Then $\mathbb{Q}$ is a field.

Definition 2.38 (Quotient map)**.** The quotient map is the map

$$\pi \colon X \to X/\sim$$
$$x \mapsto [x]$$

Lemma 2.39. *Quotient maps are surjective.*

Proof. By construction, every equivalence class $[x] \in X/\sim$ is the image of some $x \in X$, namely $\pi(x) = [x]$. □

Axiom of Choice and Its Equivalences

Definition 2.40. Let $(P, \leq)$ be a partially ordered set. Suppose $A \subset P$.

(i) $u \in P$ is an ***upper bound*** for A if $x \leq u$ for all $x \in A$.

(ii) $m \in P$ is a ***maximal element*** of P if $x \in P$ and $m \leq x$ implies $m = x$.

(iii) Similarly we define ***lower bound*** and ***minimal element***.

(iv) $C \subset P$ is called a ***chain*** if either $x \leq y$ or $y \leq x$ for all $x, y \in C$.

This terminology of partially ordered sets will often be applied to an arbitrary family of sets.

When this is done, it should be understood that the family is being regarded as a partially ordered set under the relation $\subsetneq$. Thus a maximal member of $\mathscr{A}$ is a set $M \in \mathscr{A}$ such that M is a proper subset of no other member of $\mathscr{A}$; a chain of sets is a family $\mathscr{C}$ of sets such that $A \subsetneq B$ or $B \subsetneq A$ for all $A, B \in \mathscr{C}$.

Definition 2.41. Let $\mathscr{F}$ be a family of sets. Then $\mathscr{F}$ is said to be a *family of finite character* if for each set A, we have $A \in \mathscr{F}$ if and only if each finite subset of A is in $\mathscr{F}$.

We shall need the following technical fact.

Lemma 2.42. *Let $\mathscr{F}$ be a family of finite character, and let $\mathscr{C}$ be a chain in $\mathscr{F}$. Then $\bigcup \mathscr{C} \in \mathscr{F}$.*

Proof. It suffices to show that each finite subset of $\bigcup \mathscr{C}$ is in $\mathscr{F}$. Let $F = \{x_1, \ldots, x_n\} \subset \bigcup \mathscr{C}$. Then there exist sets $C_1, \ldots, C_n \in \mathscr{C}$ such that $x_i \in C_i$ $(i = 1, \ldots, n)$. Since $\mathscr{C}$ is a chain, there exists $i_0 \in \{1, \ldots, n\}$ such that $C_i \subsetneq C_{i_0}$ for $i = 1, \ldots, n$. Then $F \subset C_{i_0} \in \mathscr{F}$. But $\mathscr{F}$ is of finite character, and so $F \in \mathscr{F}$. □

Theorem 2.43. *The following are equivalent:*

(i) Axiom of choice*: The Cartesian product of any non-empty collection of non-empty sets is non-empty.*

(ii) Tukey's lemma*: Every non-empty family of finite character has a maximal member.*

(iii) Hausdorff maximality principle*: Every non-empty partially ordered set contains a maximal chain.*

(iv) Zorn's lemma*: Every non-empty partially ordered set in which every chain has an upper bound has a maximal element.*

(v) Well-ordering principle*: Every non-empty set has a well-ordering.*

Proof. We direct the reader to Section 3 of [HS65] for the complete proof. □

Remark. It is a non-trivial result that Zorn's lemma is independent of the usual (Zermelo–Fraenkel) axioms of set theory in the sense that if the axioms of set theory are consistent, then so are these axioms together with Zorn's lemma; and if the axioms of set theory are consistent, then so are these axioms together with the negation of Zorn's lemma.

2.4 Cardinality

This section is about formalising the notion of the "size" of a set.

Finite Sets

Definition 2.44. A and B said to be ***equivalent*** (or have the same *cardinality*), denoted by $A \sim B$, if there exists a bijection $f\colon A \to B$.

Notation. For $n \in \mathbb{N}$, denote

$$\mathbb{N}_n = \{k \in \mathbb{N} \mid 1 \leq k \leq n\},$$
$$n\mathbb{N} = \{nk \mid k \in \mathbb{N}\}.$$

Definition 2.45. We say A is ***finite*** if $A \sim \mathbb{N}_n$ for some integer $n \in \mathbb{N}$, then the *cardinality* of A is $|A| = n$; A is *infinite* if A is not finite.

There are a number of "intuitively obvious" facts about finite sets. Here is an easy fact to start with:

Lemma 2.46. *Let $a_0 \in A$. There exists a bijective $f\colon A \to \mathbb{N}_{n+1}$ if and only if there exists a bijective $g\colon A \setminus \{a_0\} \to \mathbb{N}_n$.*

Proof.

$\boxed{\Longleftarrow}$ Suppose there exists a bijection

$$g\colon A \setminus \{a_0\} \to \mathbb{N}_n.$$

Define

$$f(x) = \begin{cases} g(x) & (x \in A \setminus \{a_0\}) \\ n+1 & (x = a_0) \end{cases}$$

One checks at once that f is bijective.

$\boxed{\Longrightarrow}$ Suppose there exists a bijection

$$f\colon A \to \mathbb{N}_{n+1}.$$

If $f(a_0) = n+1$, things are especially easy; in that case, the restriction $f|_{A\setminus\{a_0\}}$ is the desired bijective correspondence of $A \setminus \{a_0\}$ with $\mathbb{N}_n$.

Otherwise, let $f(a_0) = m$, and let $a_1 \in A$ be such that $f(a_1) = n+1$. Then $a_1 \neq a_0$. Define a

new function

$$h\colon A \to \mathbb{N}_{n+1}$$

by setting

$$h(x) = \begin{cases} n+1 & (x = a_0) \\ m & (x = a_1) \\ f(x) & (x \in A \setminus \{a_0, a_1\}) \end{cases}$$

It is easy to check that h is a bijection. Now we are back in the easy case; the restriction $h|_{A\setminus\{a_0\}}$ is the desired bijection of $A \setminus \{a_0\}$ with $\mathbb{N}_n$. □

From this lemma a number of useful consequences follow:

Proposition 2.47. *Suppose there exists a bijective $f\colon A \to \mathbb{N}_n$. Let $B \subsetneq A$. Then there does not exist a bijective $g\colon B \to \mathbb{N}_n$, but there exists a bijective $h\colon B \to \mathbb{N}_m$ for some $m < n$.*

Proof. Induct on n. If $n = 1$, A consists of a single element $\{a\}$, and its only proper subset B is the empty set.

Suppose the desired result holds for n; we prove it true for $n+1$. Suppose $f\colon A \to \mathbb{N}_{n+1}$ is a bijection, and $B \subsetneq A$. Choose $a_0 \in B$, $a_1 \in A \setminus B$. Apply the preceding lemma to conclude there is a bijection

$$g\colon A \setminus \{a_0\} \to \mathbb{N}_n.$$

Now $B \setminus \{a_0\}$ is a proper subset of $A \setminus \{a_0\}$. Since the result has been assumed to hold for n, we conclude the following:

(i) There exists no bijection $h\colon B \setminus \{a_0\} \to \mathbb{N}_n$.

(ii) Either $B \setminus \{a_0\} = \emptyset$, or there exists a bijection

$$k\colon B \setminus \{a_0\} \to \mathbb{N}_p \quad \text{for some } p < n.$$

The preceding lemma, combined with (i), implies that there is no bijection of B with $\mathbb{N}_{n+1}$. This is the first half of what we wanted to proved.

To prove the second half, note that if $B \setminus \{a_0\} = \emptyset$, there is a bijection of B with the set $\{1\}$; while if $B \setminus \{a_0\} \neq \emptyset$, we can apply the preceding lemma, along with (ii), to conclude that there is a bijection of B with $\mathbb{N}_{p+1}$.

In either case, there is a bijection of B with $\mathbb{N}_m$ for some $m < n+1$, as desired. □

Corollary 2.48. *If A is finite, there is no bijection of A with a proper subset of itself.*

Proof. Suppose, for a contradiction, that $B \subsetneq A$ and $f : A \to B$ is bijective.

By assumption, there is a bijection $g : A \to \mathbb{N}_n$ for some n. Then $g \circ f^{-1}$ is a bijection of B with $\mathbb{N}_n$. This contradicts the preceding result. □

Corollary 2.49. *$\mathbb{N}$ is infinite.*

Proof. The function

$$\begin{aligned} f : \mathbb{N} &\to \mathbb{N} \setminus \{1\} \\ n &\mapsto n+1 \end{aligned}$$

is a bijection of $\mathbb{N}$ with a proper subset of itself. □

Corollary 2.50. *The cardinality of a finite set A is uniquely determined by A.*

Proof. Let $m < n$. Suppose there are bijections

$$\begin{aligned} f &: A \to \mathbb{N}_n, \\ g &: A \to \mathbb{N}_m. \end{aligned}$$

Then the composite

$$g \circ f^{-1} : \mathbb{N}_n \to \mathbb{N}_m$$

is a bijection of the finite set $\mathbb{N}_n$ with a proper subset of itself, contradicting the corollary just proved. □

Corollary 2.51. *If B is a subset of the finite set A, then B is finite. If $B \subsetneq A$, then $|B| < |A|$.*

Countable and Uncountable Sets

Definition 2.52. A is ***countable*** if $A \sim \mathbb{N}$; A is *uncountable* if A is neither finite nor countable; A is *at most countable* if A is finite or countable.

Since $\mathbb{N}$ is infinite, any countable set is infinite.

Example. $\mathbb{N}$ is countable since the identity map from $\mathbb{N}$ to $\mathbb{N}$ is a bijection.

Example. $n\mathbb{N}$ is countable.

Proof. Let $f\colon \mathbb{N} \to n\mathbb{N}$ which sends $k \mapsto nk$. We need to show that f is bijective:

- For any $k_1, k_2 \in \mathbb{N}$, $nk_1 = nk_2$ implies $k_1 = k_2$ so f is injective.
- For any $x \in n\mathbb{N}$, $x = nk$ for some $k \in \mathbb{N}$, thus $\frac{x}{n} = k \in \mathbb{N}$ so f is surjective.

Hence f is bijective, so $n\mathbb{N} \sim \mathbb{N}$ and we are done. □

Example. $\mathbb{Z}$ is countable.

Proof. Consider the following arrangement of the elements of $\mathbb{Z}$ and $\mathbb{N}$:

$$\begin{aligned} \mathbb{Z}\colon &\quad 0, 1, -1, 2, -2, 3, -3, \dots \\ \mathbb{N}\colon &\quad 1, 2, 3, 4, 5, 6, 7, \dots \end{aligned}$$

In fact we can write an explicit formula for a bijection $f\colon \mathbb{N} \to \mathbb{Z}$ where

$$f(n) = \begin{cases} \dfrac{n}{2} & (n \text{ even}) \\ -\dfrac{n-1}{2} & (n \text{ odd}) \end{cases}$$

□

Remark. Any countable set can be "listed" in a sequence $a_1, a_2, \dots$ of distinct terms. This technique is particularly useful when it is difficult or not possible to deduce an explicit formula for a bijection.

Proposition 2.53. *Every infinite subset of a countable set is countable.*

Proof. Let S be the countable set. Then we can arrange the elements of S in a sequence (s_n) of distinct elements:

$$s_1, s_2, \dots$$

Suppose $E \subset S$ is infinite. The main idea is to show that we can list out the elements of E in a sequence. We now construct a sequence (n_k) as follows: Let

$$\begin{aligned} n_1 &= \min\{i \mid s_i \in E\} \\ n_2 &= \min\{i \mid s_i \in E, i > n_1\} \\ &\vdots \\ n_k &= \min\{i \mid s_i \in E, i > n_{k-1}\}. \end{aligned}$$

Then

$$E = \{s_{n_1}, s_{n_2}, \dots\},$$

where we note that the function $f(k) = s_{n_k}$ $(k = 1, 2, \dots)$ is bijective. Hence $E \sim \mathbb{N}$, as desired. □

Remark. This shows that countable sets represent the "smallest" infinity: No uncountable set can be a subset of a countable set.

Proposition 2.54. *The countable union of countable sets is countable.*

Proof. Let $\{A_n \mid n \in \mathbb{N}\}$ be a fanily of countable sets; clearly this is a countable collection of sets (indexed by $\mathbb{N}$). Then we want to show that the union

$$S = \bigcup_{n=1}^{\infty} A_n$$

is countable.

Since every set A_n is countable, we can list its elements in a sequence (a_{nk}) $(k = 1, 2, 3, \dots)$. Arrange the elements of all the sets in $\{A_n\}$ in the form of an infinite array, containing all elements of S, where the elements of A_n form the n-th row.

$$\begin{array}{ll} A_1\colon & \cancel{a_{11}} \quad \cancel{a_{12}} \quad \cancel{a_{13}} \quad \cancel{a_{14}} \quad \cdots \\ A_2\colon & \cancel{a_{21}} \quad \cancel{a_{22}} \quad \cancel{a_{23}} \quad \cancel{a_{24}} \quad \cdots \\ A_3\colon & \cancel{a_{31}} \quad \cancel{a_{32}} \quad \cancel{a_{33}} \quad \cancel{a_{34}} \quad \cdots \\ A_4\colon & \cancel{a_{41}} \quad \cancel{a_{42}} \quad \cancel{a_{43}} \quad \cancel{a_{44}} \quad \cdots \\ \vdots & \end{array}$$

We then zigzag our way through the array, and arrange these elements in a sequence

$$a_{11}, a_{21}, a_{12}, a_{31}, a_{22}, a_{13}, a_{41}, a_{32}, a_{23}, a_{14}, \dots$$

thus S is countable, and we are almost done!

A small problem is that if any two of the sets A_n have elements in common, these will appear more than once in the above sequence. Then we take a subset $T \subset S$, where every element only appears once. Note that T is an infinite subset, since $A_1 \subset T$ is infinite. Then since T is an infinite subset of a countable set S, by Proposition 2.53, T is countable. □

Remark. If we were to instead start by going down by the first row of the above array, then we would not get to the second row (and beyond); all that would show is the first row is countable. Instead, we wind our way through diagonally, ensuring that we hit every number of the array.

Corollary 2.55. *Suppose A is an indexing set that is at most countable. Let $\{B_\alpha \mid \alpha \in A\}$ be a family of sets that are at most countable. Then the union*

$$\bigcup_{\alpha \in A} B_\alpha$$

is at most coutable.

Proposition 2.56. *Let A be a countable set. For $n \in \mathbb{N}$, let*

$$B_n = \{(a_1, \ldots, a_n) \mid a_i \in A\}.$$

Then B_n is countable.

Proof. We prove by induction on n. That B_1 is countable is evident, since $B_1 = A$.

Now suppose B_{n-1} is countable. The elements of B_n are of the form

$$(b,a) \quad (b \in B_{n-1}, a \in A)$$

For every fixed b, the set of ordered pairs (b,a) is equivalent to A, and hence countable. Thus B_n is a union of countable sets. By Proposition 2.54, B_n is countable. □

Corollary 2.57. *$\mathbb{Q}$ is countable.*

Proof. Note that every $x \in \mathbb{Q}$ is of the form $\frac{b}{a}$, where $a, b \in \mathbb{Z}$. By the previous result, taking $n = 2$, the set of pairs (a,b) and therefore the set of fractions $\frac{b}{a}$ is countable. □

That not all infinite sets are, however, countable, is shown by the next result.

Proposition 2.58. *Let A be the set of all sequences whose elements are the digits 0 and 1. Then A is uncountable.*

Proof. Let $E \subset A$ be countable, consisting of the sequences $s_1, s_2, s_3, \ldots$.

We construct a new sequence s as follows:

$$n\text{-th digit of } s = \begin{cases} 0 & \text{if } n\text{-th digit in } s_n \text{ is } 1, \\ 1 & \text{if } n\text{-th digit in } s_n \text{ is } 0. \end{cases}$$

Then the sequence s differs from every member of E in at least one place, so $s \notin E$. But clearly $s \in A$; hence $E \subsetneq A$.

We have shown that every countable subset of A is a proper subset of A. It follows that A is uncountable (for otherwise A would be a proper subset of A, which is absurd). □

Remark. The idea of the above proof is called *Cantor's diagonal process*, first used by Cantor. This is because if elements of the sequences $s_1, s_2, s_3, \ldots$ are listed out in an array, it is the elements on the diagonal which are involved in the construction of the new sequence.

Corollary 2.59. *$\mathbb{R}$ is uncountable.*

Proof. This follows from the binary representation of the real numbers. □

Theorem 2.60 (Cantor's theorem)**.** *For any set A, we have $A \not\sim \mathscr{P}(A)$.*

Proof. Suppose otherwise, for a contradiction, that $A \sim \mathscr{P}(A)$. Then there exists a bijection $f : A \to \mathscr{P}(A)$. Then for each $x \in A$, $f(x)$ is a subset of A. Now consider the "anti-diagonal" set

$$B = \{x \in A \mid x \notin f(x)\}.$$

That is, B is the subset of A containing all $x \in A$ such that x is not in the set $f(x)$. Since $B \subset A$, we have $B \in \mathscr{P}(A)$. Since f is bijective (in particular surjective), there exists $x \in A$ such that $f(x) = B$. Now there are two cases: (i) $x \in B$, or (ii) $x \notin B$.

(i) If $x \in B$, then by definition of the set B it must be the case that $x \notin f(x)$. But since $f(x) = B$, we then have $x \notin D$. This is absurd since we cannot have $x \in B$ and $x \notin B$ simultaneously.

(ii) If $x \notin B$, by definition of the set B, this implies that $x \in f(x)$. But $f(x) = B$. So we have $x \in B$ and $x \notin B$, which is again absurd.

In either case, we have reached a contradiction. Hence there cannot exist a surjective (and thus bijective) function $A \to \mathscr{P}(A)$. □

Exercises

Exercise 2.1. Prove that the statement $\forall x \in \emptyset, P(x)$ is vacuously true.

Solution. Let S be the embedding set. The statement $\forall x \in \emptyset, P(x)$ means

$$\forall x \in S, \quad x \in \emptyset \implies P(x).$$

But $x \in \emptyset$ is always false, by the definition of empty set. Hence the statement is always true, regardless of x. □

Exercise 2.2. Prove that for any set $A \subset S$, $\emptyset \subset A$ and $A \subset A$.

Solution. Let $A \subset S$. Let $x \in \emptyset$, then $x \in \emptyset \implies x \in A$ is vacuously true, so $\emptyset \subset A$.

Likewise, let $x \in A$, then $x \in A \implies x \in A$ is always true, so $A \subset A$. □

Exercise 2.3. Let A be the set of all complex polynomials in n variables. Given a subset $T \subset A$, define the *zeros* of T as the set

$$Z(T) = \{P = (a_1, \ldots, a_n) \mid f(P) = 0 \text{ for all } f \in T\}.$$

$Y \subset \mathbb{C}^n$ is called an *algebraic* set if there exists $T \subset A$ such that $Y = Z(T)$.
Prove that the union of two algebraic sets is an algebraic set.

Solution. We would like to consider $T = \{f_1, f_2, \ldots\}$ expressed as indexed sets $T = \{f_i\}$. Then $Z(T)$ can also be expressed as $\{P \mid \forall i, f_i(P) = 0\}$.

Suppose that we have two algebraic sets X and Y. Let $X = Z(S)$, $Y = Z(T)$ where S, T are subsets of A (basically, they are certain sets of polynomials). Then

$$X = \{P \mid \forall f \in S, f(P) = 0\}$$

$$Y = \{P \mid \forall g \in T, g(P) = 0\}$$

We imagine that for $P \in X \cap Y$, we have $f(P) = 0$ or $g(P) = 0$. Hence we consider the set of polynomials

$$U = \{f \cdot g \mid f \in S, g \in T\}$$

For any $P \in X \cup Y$ and for any $fg \in U$ where $f \in S$ and $f \in g$, either $f(P) = 0$ or $g(P) = 0$, hence $fg(P) = 0$ and thus $P \in Z(U)$.

On the other hand if $P \in Z(U)$, suppose otherwise that P is not in $X \cup Y$, then P is neither in X nor in Y. This means that there exists $f \in S, g \in T$ such that $f(P) \neq 0$ and $g(P) \neq 0$, hence

$fg(P) \neq 0$. This is a contradiction as $P \in Z(U)$ implies $fg(P) = 0$. Hence we have $X \cup Y = Z(U)$ and thus $X \cup Y$ is an algebraic set.

Now the other direction is simpler and can actually be generalised: The intersection of arbitrarily many algebraic sets is algebraic.

The basic result is that if $X = Z(S)$ and $Y = Z(T)$ then $X \cap Y = Z(S \cup T)$. □

Exercise 2.4 (Complex numbers)**.** Let $\mathbb{R}[x]$ denote the set of real polynomials. Define

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

where

$$f(x) \sim g(x) \iff x^2+1 \text{ divides } f(x) - g(x).$$

The complex number $a + bi$ is defined to be the equivalence class of $a + bx$.

(a) Define the sum and product of two complex numbers and show that such definitions are well-defined.

(b) Define the reciprocal of a complex number.

Exercise 2.5 ([Rud76] 2.2)**.** We say $z \in \mathbb{C}$ is *algebraic* if there exist integers $a_0, \ldots, a_n$, not all zero, such that

$$a_0 z^n + a_1 z^{n-1} + \cdots + a_{n-1} z + a_n = 0.$$

Prove that the set of all algebraic numbers is countable. *Hint*: For every positive integer N there are only finitely many equations with

$$n + |a_0| + |a_1| + \cdots + |a_n| = N.$$

Solution. Following the hint, let A_N be the set of numbers z that satisfy $a_0 z^n + a_1 z^{n-1} + \cdots + a_{n-1} z + a_n = 0$, for some coefficients $a_0, \ldots, a_n$ which satisfy

$$n + |a_0| + |a_1| + \cdots + |a_n| = N.$$

By the fundamental theorem of algebra, $a_0 z^n + a_1 z^{n-1} + \cdots + a_{n-1} z + a_n = 0$ has at most n solutions, so each A_N is finite. Hence the set of algebraic numbers, which is the union

$$\bigcup_{N=2}^{\infty} A_N$$

is at most countable. Since all rational numbers are algebraic, it follows that the set of algebraic numbers is exactly countable. □

Exercise 2.6 ([Rud76] 2.3)**.** Prove that there exist real numbers which are not algebraic.

Solution. By the previous exercise, the set of real algebraic numbers is countable. If every real number were algebraic, the entire set of real numbers would be countable, a contradiction. □

Exercise 2.7 ([Rud76] 2.4)**.** Is the set of irrational real numbers countable?

Solution. No. If $\mathbb{R} \setminus \mathbb{Q}$ were countable, $\mathbb{R} = \mathbb{Q} \cup (\mathbb{R} \setminus \mathbb{Q})$ would be countable, which is clearly false. □

Exercise 2.8. Negate surjectivity.

Solution. If $f : X \to Y$ is not surjective, then it means that there exists $y \in Y$ not in the image of X, i.e. for all x in X we have $f(x) \neq y$.

$$\begin{aligned}
&\neg \forall y \in Y, \exists x \in X, f(x) = y \\
\equiv &\exists y \in Y, \neg(\exists x \in X, f(x) = y) \\
\equiv &\exists y \in Y, \forall x \in X, \neg(f(x) = y) \\
\equiv &\exists y \in Y, \forall x \in X, f(x) \neq y
\end{aligned}$$

□

Exercise 2.9 (NUS MA1100T AY23/24)**.** Define a relation $\sim$ on $\mathbb{R}$ such that $x \sim y$ if $x - y \in \mathbb{Q}$.

(i) Prove that $\sim$ is an equivalence relation on $\mathbb{R}$.

(ii) Prove that the quotient set $\mathbb{R}/\sim$ is infinite.

Solution.

(i) For all $x \in \mathbb{R}$, $x - x = 0 \in \mathbb{Q}$. Hence $x \sim x$ for all $x \in \mathbb{R}$, so $\sim$ is reflexive.

Suppose $x, y \in \mathbb{R}$ are such that $x \sim y$. Then $x - y \in \mathbb{Q}$, so $y - x = -(x - y) \in \mathbb{Q}$ implies $y \sim x$. Hence $\sim$ is symmetric.

Suppose $x, y, z \in \mathbb{R}$ are such that $x \sim y$ and $y \sim z$. Then $x - y, y - z \in \mathbb{Q}$, so $x - z = (x - y) + (y - z) \in \mathbb{Q}$ implies $x \sim z$. Hence $\sim$ is transitive.

(ii)

Idea. To show a set $(\mathbb{R}/\sim)$ is infinite, it suffices to show that one of its subsets (a set consisting of equivalence classes) is infinite.

Consider the set $S = \left\{ [\sqrt{2}k] \mid k \in \mathbb{Z} \right\}$. We will show that S is infinite. (Since $S \subset \mathbb{R}/\sim$, this will imply that $\mathbb{R}/\sim$ is infinite.)

Define the map

$$\begin{aligned} f\colon \mathbb{Z} &\to S \\ k &\mapsto [\sqrt{2}k] \end{aligned}$$

We shall show that f is injective. (Since $\mathbb{Z}$ is infinite, this will imply that S is infinite.)

Suppose $[\sqrt{2}k] = [\sqrt{2}k']$. Then $\sqrt{2}k \sim \sqrt{2}k'$. Thus $\sqrt{2}k - \sqrt{2}k' = (k-k')\sqrt{2} \in \mathbb{Q}$.

Since $k - k' \in \mathbb{Z}$ and $\sqrt{2}$ is irrational, we must have $k - k' = 0$. Hence $k = k'$, so f is injective.

□

Exercise 2.10 (NUS MA1100T AY23/24)**.** Suppose J is a non-empty indexing set. Let $(A_j)_{j\in J}$ be a family of sets, and define their *disjoint union* as

$$\bigsqcup_{j\in J} A_j = \{(j,a) \mid j \in J, a \in A_j\}.$$

For each $j \in J$, define the function $i_j\colon A_j \mapsto \bigsqcup_{j\in J} A_j$ by $a \mapsto (j,a)$. Define also the family of functions $(f_j\colon A_j \to X)_{j\in J}$ (consisting of one such f_j for each $j \in J$). Prove that there exists a unique function $f\colon \bigsqcup_{j\in J} A_j \to X$ such that $f_j = f \circ i_j$ for each $j \in J$.

Solution. Define f by $(j,a) \mapsto f_j(a)$. We first prove f is well-defined.

$f(j,a)$ always exists, because the family of functions has one f_j for every $j \in J$.

Then, Suppose $(j_1,a_1) = (j_2,a_2)$. Then, $j_1 = j_2$, so $f_{j_1} = f_{j_2}$, and $a_1 = a_2$.

Since all the f_j's are well defined functions, and $a_1 = a_2$, we have

$$f(j_1,a_1) = f_{j_1}(a_1) = f_{j_1}(a_2) = f_{j_2}(a_2) = f(j_2,a_2)$$

so f is well-defined.

Then, for each $j \in J$, we have that

$$f \circ i_j(a) = f(i_j(a)) = f(j,a) = f_j(a)$$

for all $a \in A_j$, so $f_j = f \circ i_j$.

To show f is unique, suppose $g : \bigsqcup_{j\in J} A_j \to X$ is a function such that $f_j = g \circ i_j$ for each $j \in J$.

Note that i_j is onto, as for each $(j,a) \in \bigsqcup_{j\in J} A_j$, $i_j(a) = (j,a)$.

Then we have the following equalities that hold for all $(j,a) \in \bigsqcup_{j\in J} A_j$:

$$\begin{aligned} g(j,a) &= g(i_j(a)) \quad \text{[because } i_j \text{ is onto]} \\ &= f_j(a) \\ &= f(i_j(a)) \\ &= f(j,a) \end{aligned}$$

so $g = f$ as desired. □

Exercise 2.11 (NUS MA1100T AY23/24)**.** Suppose $m,n \in \mathbb{N}$. We attempt to define a function $f : [mn] \to [m] \times [n]$ by

$$f(R_{mn}(a)) = (R_m(a), R_n(a)).$$

Prove that f is well-defined.

Solution. Suppose $c \in [mn]$. Since the R_b function is onto for all $b \in \mathbb{N}^+$, there exists some integer a such that $c = R_{mn}(a)$. Hence, $R_m(a)$ and $R_n(a)$ exist, so $f(c)$ exists.

Suppose now that $R_{mn}(a_1) = c = R_{mn}(a_2)$.

Then, $a_1 = q_1(mn) + c$ and $a_2 = q_2(mn) + c$, for some $q_1, q_2 \in \mathbb{Z}$.

Then, $a_1 - a_2 = (q_1 - q_2)mn$, so $m \mid a_1 - a_2$. Hence, $R_m(a_1) = R_m(a_2)$.

Likewise, $n \mid a_1 - a_2$, so $R_n(a_1) = R_n(a_2)$.

Hence, $(R_m(a_1), R_n(a_1)) = (R_m(a_2), R_n(a_2))$, and $f(c)$ has a unique value. □

Exercise 2.12 (NUS MA1100T AY23/24)**.** Let C be the set of functions from $\mathbb{N}$ to $\{0,1\}$ that are *eventually constant*, that is, for each $f \in C$, there exists $n \in \mathbb{N}$ such that for all $m > n$, $f(m) = f(n)$. For each $f \in C$, define the set S_f by

$$S_f = \{n \in \mathbb{N} : (\forall m > n)(f(m) = f(n))\}.$$

Define the function $F : C \to \{0,1\}^{<\mathbb{N}}$ by

$$F(f) = (f(0), f(1), \ldots, f(\min(S_f))).$$

(i) Prove that F is one-to-one.

(ii) Prove that C is countable.

Solution.

(i) Suppose $F(f_1) = F(f_2)$. Then, by the definition of F, we have

$$(f_1(0), f_1(1), \ldots, f_1(\min(S_{f_1}))) = (f_2(0), f_2(1), \ldots, f_2(\min(S_{f_2}))).$$

For the two sequences to be equal, they must have equal length. Hence, $\min(S_{f_1}) = \min(S_{f_2})$.

Let $\min(S_{f_1}) = \min(S_{f_2}) = k$. For the two sequences to be equal, they are also termwise equal. Hence, for all $m \leq k$, $f_1(m) = f_2(m)$.

By the definition of S_f, for all $m > k$, $f(m) = f(k)$. Hence, $f_1(m) = f_1(k) = f_2(k) = f_2(m)$.

Hence $f_1(m) = f_2(m)$ for all $m \in \mathbb{N}$, so $f_1 = f_2$, as desired.

(ii) Since $\{0,1\}$ is finite, $\{0,1\}^{<\mathbb{N}}$ is countably infinite.

Hence, there exists a bijection g from $\{0,1\}^{<\mathbb{N}}$ to $\mathbb{N}$.

Then, $g \circ F$ is an injection from C to $\mathbb{N}$, so C is countable.

□

Exercise 2.13 (NUS MA1100T AY23/24)**.** Prove that for every pair of real numbers $q < r$, there exists an irrational number that is strictly between them.

Solution. Since $\mathbb{Q}$ is dense in $\mathbb{R}$, there is a rational number x strictly between q and r.

Then, by the same argument, there is a rational number y strictly between x and r.

Then we know there is an irrational number strictly between two rational numbers, so we are done, as there is an irrational z such that $q < x < z < y < r$. □

Exercise 2.14 (NUS MA1100T AY21/22)**.** Let X be any set such that $\emptyset \in X$ and such that for any $x \in X$, one has $\{x\} \in X$. The sequence $A_1, A_2, \ldots$ of elements of X is defined recursively as follows:

$$A_1 = \emptyset$$
$$A_{n+1} = \{A_n\} \quad (n \in \mathbb{N}).$$

Show thay for any $i, j \in \mathbb{N}$ with $i \neq j$, one has $A_i \neq A_j$.

Solution. Let $P(n)$ be the proposition that for any $i, j \in \mathbb{N}$ with $i, j \leq n$ and $i \neq j$, $A_i \neq A_j$. We shall prove $P(n)$ for all n via induction.

(Base case) If $n = 1$, then the proposition is vacuously true as i and j must both be equal to 1. If $n = 2$ then $i = 1$ and $j = 2$ without loss of generality. Then $A_1 = \emptyset$ and $A_2 = \{\emptyset\}$ so $A_1 \neq A_2$. This proves the base case.

(Inductive Step) Suppose that $P(n)$ is true for some positive integer $n \geq 2$. We will prove $P(n+1)$ is true. Let $i, j \leq n+1$. Note if $i, j \leq n$ then by assumption, $A_i \neq A_j$. Hence, i or j must be $n+1$. Without loss of generality, let $i = n+1$. Since $j = n+1 \Rightarrow i = j$, we must have $j \leq n$. If $j = 1$ then $A_{n+1} \neq A_1$ since A_1 is empty and A_{n+1} is not. Otherwise, $j > 1$. Suppose toward a contradiction that $A_{n+1} = A_j$. Since A_n and A_{j-1} are the only elements of the sets A_{n+1} and A_j respectively, $A_n = A_{j-1}$. But $n, j-1 \in \mathbb{N}$ and $n, j-1 \leq n$ with $n \neq j-1$ therefore by assumption $A_n \neq A_{j-1}$. This is a contradiction, hence, $A_{n+1} \neq Aj$. We have shown that $P(n+1)$ is true which completes the inductive step.

Now for any $i, j \in \mathbb{N}$, take $n = \max i, j$. Since $i, j \leq n$ and $i \neq j$, $P(n)$ witnesses $A_i \neq A_j$ as desired. □

Exercise 2.15. Let X, Y be sets and let $f\colon X \to Y$ be a map. Prove or disprove: f is injective if and only if for any set T, the "post-composition with f" map

$$\begin{aligned} \Phi_T\colon \mathrm{Maps}(T,X) &\to \mathrm{Maps}(T,Y) \\ \phi &\mapsto f \circ \phi \end{aligned}$$

is injective.

Solution. True.

$\boxed{\Longrightarrow}$ Suppose f is injective.

Let $\phi_1, \phi_2 \in \mathrm{Maps}(T,X)$ such that $\Phi_T(\phi_1) = \Phi(\phi_2)$. This implies $f \circ \phi_1 = f \circ \phi_2$. Hence, for all $t \in T$, $f(\phi_1(t)) = f(\phi_2(t))$.

Since f is injective, $\phi_1(t) = \phi_2(t)$ for all $t \in T$. Therefore, ϕ_1 and ϕ_2 are the same function. This proves that Φ_T is injective.

$\boxed{\Longleftarrow}$ Suppose Φ_T is injective for any set T.

Pick $T = \{0\}$, then Φ_T is injective. For all $x, y \in X$ such that $f(x) = f(y)$, choose functions $\phi_x, \phi_y \in$ Maps(T, X) such that $\phi_x(0) = x$ and $\phi_y(0) = y$. Then $f(x) = f(y) \Rightarrow f(\phi_x(t)) = f(\phi_y(t))$ for all $t \in T = \{0\}$, i.e. $f \circ \phi_x = f \circ \phi_y$. But then $\Phi_T(\phi_x) = f \circ \phi_x = f \circ \phi_y = \Phi_T(\phi_y)$. Since Φ_T is injective, it follows that $\phi_x = \phi_y$. Hence, $x = \phi_x(0) = \phi_y(0) = y$. We conclude that f is injective. □

Exercise 2.16. Let X, Y be sets and let $f : X \to Y$ be a map. Prove or disprove: f is surjective if and only if for any set T, the "pre-composition with f" map

$$\begin{aligned} \Psi_T\colon \mathrm{Maps}(Y,T) &\to \mathrm{Maps}(X,T) \\ \psi &\mapsto \psi \circ f \end{aligned}$$

is surjective.

Solution. False.

Consider the sets $X = \{1,2\}$, $Y = \{3\}$. We will prove that for set $T = X$, Ψ_T is not surjective.

Define the function $f : X \to Y$ by $f(x) = 3$. Since $f(X) = \{3\} = Y$, so f is surjective. For any $\psi \in \text{Maps}(Y,T)$, note that $\psi(f(1)) = \psi(3) = \psi(f(2))$ but $1 \neq 2$. Hence, $\psi \circ f$ is not injective, and is therefore not the identity function, id_X. This proves that $\text{id}_X \notin \text{Range}(\Psi_T)$. Since, $T = X$, $\text{id}_X \in \text{Maps}(X,T)$ hence, Ψ_T is not surjective. We conclude that the forward direction does not hold, hence the statement is false. □

Exercise 2.17 (NUS MA1100 AY24/25)**.** For each $n \in \mathbb{Z}^+$, let

$$A_n = \left(1 - \frac{1}{n}, n\right) = \left\{x \in \mathbb{R} \mid 1 - \frac{1}{n} < x < n\right\}.$$

Find the sets

$$\bigcup_{n=1}^{\infty} A_n \qquad \text{and} \qquad \bigcap_{n=1}^{\infty} A_n$$

Justify you answers.

Solution. We claim that

$$\bigcup_{n=1}^{\infty} A_n = (0,\infty).$$

To prove $\subseteq$, first let $x \in \bigcup_{n=1}^{\infty} A_n$. Then, there exists $n \in \mathbb{Z}^+$ such that

$$1 - \frac{1}{n} < x < n$$

When $n = 1$, we have $0 < x < 1$. As $n \to \infty$, we have $x > 1$. As such, $0 < x$ which follows that $x \in (0,\infty)$

To prove the reverse inclusion $\supseteq$, let $x \in (0,\infty)$. Then for some $n \in \mathbb{Z}^+$, we can pick $n > \max\{x, \frac{1}{x}\}$ such that we have

$$x < n \qquad \text{or} \qquad \frac{1}{n} > x \implies -\frac{1}{n} < -x \implies 1 - \frac{1}{n} < 1 - x < x.$$

Hence,

$$1 - \frac{1}{n} < x < n \implies x \in \bigcup_{n=1}^{\infty} A_n.$$

We then claim that

$$\bigcap_{n=1}^{\infty} A_n = \emptyset.$$

Suppose for the sake of contradiction that the intersection is non empty. Then, there exists $x \in A_n$ such that

$$1 - \frac{1}{n} < x < n.$$

Since n is increasing, then $x < n$ is trivially satisfied for fixed x, since $n \to \infty$. As $n \to \infty$, the lower bound will be 1. As such $x \in (1, \infty)$. We have $A_1 = (0,1)$ which is disjoint with $(1,\infty)$. Therefore, $x \notin A_1$ which is a contradiction. □

Exercise 2.18 (NUS MA1100 AY24/25)**.** Let X and Y be nonempty sets and let $f : X \to Y$ be a function. Prove that if f is surjective, then X has a subset Z such that the function $h : Z \to Y$ defined by

$$h(x) = f(x) \quad \text{for all } x \in Z$$

is a bijection.

Solution. Since f is surjective, then for all $y \in Y$, there exists $x \in X$ such that $f(x) = y$. Define the set

$$Z = \{x \in X \mid y \in Y\} \subseteq X \quad \text{so} \quad h(x) = f(x) = y.$$

Suppose $h(x_1) = h(x_2)$, then $f(x_1) = f(x_2)$. which implies $y_1 = y_2$. Since f is surjective, then $x_1 = x_2$ which also shows that $h(x)$ is injective.

We then prove that h is surjective. Let $y \in Y$. Then, by the definition of Z, there exists $x \in Z$ such that $h(x) = y$. This shows that $h(x)$ is surjective. Since $h(x)$ is both injective and surjective, then h is a bijection. □

Exercise 2.19 (NUS MA1100 AY24/25)**.** Let X and Y be two nonempty sets, and let $f : X \to Y$ be a surjective function. Let $\sim$ be the relation on X defined by, for all $x, y \in X$,

$$x \sim y \quad \text{if and only if} \quad f(x) = f(y)$$

(i) Prove that $\sim$ is an equivalence relation.

(ii) Prove that $X/\sim$ is equinumerous with Y.

Solution.

(i) For all $x \in X$, we have $x \sim x \iff f(x) = f(x)$. Hence $\sim$ is reflexive.

Suppose $x \sim y$. Then $x \sim y \iff f(x) = f(y) \iff f(y) = f(x) \iff y \sim x$. Hence $\sim$ is symmetric.

Suppose $x \sim y$ and $y \sim z$. Then $f(x) = f(y)$ and $f(y) = f(z)$, so $f(x) = f(z) \iff x \sim z$. Hence $\sim$ is transitive.

(ii) Define the map

$$\begin{aligned} \phi : X/\sim &\to Y \\ [x] &\mapsto f(x). \end{aligned}$$

We claim that ϕ is a bijection.

- Suppose $[x] = [x']$. Then $x \sim x' \iff f(x) = f(x') \implies \phi([x]) = \phi([x'])$. This shows ϕ is well-defined.
- Suppose $\phi([x]) = \phi([x'])$. Then $f(x) = f(x') \implies x \sim x' \implies [x] = [x']$. This shows ϕ is injective.
- Since it is given that f is surjective, then for every $y \in Y$, there exist $x \in X$ such that $f(x) = y$. Then $\phi([x]) = f(x) = y$, so ϕ is surjective.

Hence $X/\sim$ is equinumerous with Y.

□

II
Linear Algebra

3 Vector Spaces

3.1 Definition of Vector Space

Let $\mathbf{F}$ denote a field, which can mean either $\mathbb{R}$ or $\mathbb{C}$.

Definition 3.1 (Vector space). V is a ***vector space*** over $\mathbf{F}$ if the following properties hold:

(i) Addition is commutative: $u+v=v+u$ for all $u,v\in V$

(ii) Addition is associative: $(u+v)+w=u+(v+w)$ for all $u,v,w\in V$
Multiplication is associative: $(ab)v=a(bv)$ for all $v\in V$, $a,b\in\mathbf{F}$

(iii) Additive identity: there exists $\mathbf{0}\in V$ such that $v+\mathbf{0}=v$ for all $v\in V$

(iv) Additive inverse: for every $v\in V$, there exists $w\in V$ such that $v+w=\mathbf{0}$

(v) Multiplicative identity: $1v=v$ for all $v\in V$

(vi) Distributive properties: $a(u+v)=au+av$ and $(a+b)v=av+bv$ for all $a,b,\in\mathbf{F}$ and $u,v\in V$

Notation. For the rest of this text, V denotes a vector space over $\mathbf{F}$.

Elements of a vector space are called *vectors* or *points*.

Remark. The scalar multiplication in a vector space depends on $\mathbf{F}$. Thus when we need to be precise, we will say that V is a vector space *over* $\mathbf{F}$.

A vector space over $\mathbb{R}$ is called a *real vector space*; a vector space over $\mathbb{C}$ is called a *complex vector space*.

Lemma 3.2 (Uniqueness of additive identity). *A vector space has a unique additive identity.*

Proof. Suppose that $\mathbf{0}$ and $\mathbf{0}'$ are additive identities of V. Then

$$\mathbf{0}' = \mathbf{0}' + \mathbf{0} = \mathbf{0} + \mathbf{0}' = \mathbf{0}$$

where the first equality holds because $\mathbf{0}$ is an additive identity, the second equality comes from commutativity, and the third equality holds because $\mathbf{0}'$ is an additive identity. □

Lemma 3.3 (Uniqueness of additive inverse)**.** *Every element in a vector space has a unique additive inverse.*

Proof. Let $v \in V$. Suppose w and w' are additive inverses of v. Then

$$w = w + \mathbf{0} = w + (v + w') = (w + v) + w' = \mathbf{0} + w' = w'.$$

□

Because additive inverses are unique, the following notation now makes sense.

Notation. Let $v, w \in V$. Then $-v$ denotes the additive inverse of v, and define $w - v$ to be $w + (-v)$.

We now prove some seemingly trivial facts.

Lemma 3.4.

(i) For every $v \in V$, $0v = \mathbf{0}$.

(ii) For every $a \in \mathbf{F}$, $a\mathbf{0} = \mathbf{0}$.

(iii) For every $v \in V$, $(-1)v = -v$.

Proof.

(i) Let $v \in V$,

$$0v = (0 + 0)v = 0v + 0v.$$

Adding the additive inverse of $0v$ to both sides of the equation gives $\mathbf{0} = 0v$.

(ii) Let $a \in \mathbf{F}$,

$$a\mathbf{0} = a(\mathbf{0} + \mathbf{0}) = a\mathbf{0} + a\mathbf{0}.$$

Adding the additive inverse of $a\mathbf{0}$ to both sides of the equation gives $\mathbf{0} = a\mathbf{0}$.

(iii) Let $v \in V$,

$$v + (-1)v = 1v + (-1)v = (1 + (-1))v = 0v = \mathbf{0}.$$

Since $v + (-1)v = \mathbf{0}$, $(-1)v$ is the additive inverse of v.

□

Example (n-tuple space)**.** Let $\mathbf{F}^n$ be the set of n-tuples whose elements belong to $\mathbf{F}$:

$$\mathbf{F}^n := \{(x_1, \dots, x_n) \mid x_i \in \mathbf{F}\}$$

For $x = (x_1, \dots, x_n) \in \mathbf{F}^n$ and $i = 1, \dots, n$, we say that x_i is the i-th *coordinate* of x. Define addition and scalar multiplication on $\mathbf{F}^n$ as

$$(x_1, \dots, x_n) + (y_1, \dots, y_n) = (x_1 + y_1, \dots, x_n + y_n)$$
$$\lambda(x_1, \dots, x_n) = (\lambda x_1, \dots, \lambda x_n)$$

Then $\mathbf{F}^n$ is a vector space over $\mathbf{F}$.

Example. Let $\mathbf{F}^\infty$ be the set of all sequences of elements of $\mathbf{F}$:

$$\mathbf{F}^\infty := \{(x_1, x_2, \dots) \mid x_i \in \mathbf{F}\}$$

Define addition and scalar multiplication on $\mathbf{F}^\infty$ as

$$(x_1, x_2, \dots) + (y_1, y_2, \dots) = (x_1 + y_1, x_2 + y_2, \dots)$$
$$\lambda(x_1, x_2, \dots) = (\lambda x_1, \lambda x_2, \dots)$$

Then $\mathbf{F}^\infty$ is a vector space over $\mathbf{F}$, where the additive identity is $\mathbf{0} = (0, 0, \dots)$.

Example (Space of functions from a set to a field)**.** If S is a non-empty set, $\mathbf{F}^S := \{f \mid f \colon S \to \mathbf{F}\}$. Define addition and scalar multiplication on $\mathbf{F}^S$ as

$$(f + g)(x) = f(x) + g(x) \quad (x \in S)$$
$$(\lambda f)(x) = \lambda f(x) \quad (x \in S)$$

for all $f, g \in \mathbf{F}^S$, $\lambda \in \mathbf{F}$. Then $\mathbf{F}^S$ is a vector space over $\mathbf{F}$, where the additive identity of $\mathbf{F}^S$ is the function $0 \colon S \to \mathbf{F}$ defined as

$$0(x) = 0 \quad (\forall x \in S)$$

and for $f \in \mathbf{F}^S$, additive inverse of f is the function $-f \colon S \to \mathbf{F}$ defined as

$$(-f)(x) = -f(x) \quad (\forall x \in S)$$

Remark. $\mathbf{F}^n$ and $\mathbf{F}^\infty$ are special cases of the vector space $\mathbf{F}^S$; think of $\mathbf{F}^n$ as $\mathbf{F}^{\{1,2,\dots,n\}}$, and $\mathbf{F}^\infty$ as $\mathbf{F}^{\{1,2,\dots\}}$.

Example (Complexification)**.** Suppose V is a real vector space. The *complexifcation* of V is $V_{\mathbb{C}} := V \times V$. An element of $V_{\mathbb{C}}$ is an ordered pair (u,v), where $u,v \in V$, which we write as $u+iv$.

- Addition on $V_{\mathbb{C}}$ is defined as

$$(u_1 + iv_1) + (u_2 + iv_2) = (u_1 + u_2) + i(v_1 + v_2)$$

  for all $u_1, v_2, u_2, v_2 \in V$.

- Complex scalar multiplication on $V_{\mathbb{C}}$ is defined as

$$(a+bi)(u+iv) = (au - bv) + i(av + bu)$$

  for all $a,b \in \mathbb{R}$ and all $u,v \in V$.

Then $V_{\mathbb{C}}$ is a (complex) vector space.

3.2 Subspaces

Whenever we have a mathematical object with some structure, we want to consider subsets that also have the same structure.

Definition 3.5 (Subspace)**.** We say $U \subset V$ is a ***subspace*** of V, denoted as $U \leq V$, if U is also a vector space (with the same addition and scalar multiplication as on V).

The sets $\{\mathbf{0}\}$ and V are always subspaces of V. The subspace $\{\mathbf{0}\}$ is called the *zero subspace* or *trivial subspace*. Subspaces other than V are called *proper subspaces*.

The following result is useful in determining whether a given subset of V is a subspace of V.

Lemma 3.6 (Subspace test)**.** *Suppose $U \subset V$. Then $U \leq V$ if and only if U satisfies the following conditions:*

(i) $\mathbf{0} \in U$. *(additive identity)*

(ii) $u + w \in U$ *for all* $u, w \in U$. *(closed under addition)*

(iii) $\lambda u \in U$ *for all* $\lambda \in \mathbf{F}$, $u \in U$. *(closed under scalar multiplication)*

Proof.

$\boxed{\Longrightarrow}$ If $U \leq V$, then U satisfies the three conditions above by the definition of vector space.

$\boxed{\Longleftarrow}$ Suppose U satisfies the three conditions above. (i) ensures that the additive identity of V is in U. (ii) ensures that addition makes sense on U. (iii) ensures that scalar multiplication makes sense on U.

If $u \in U$, then $-u = (-1)u \in U$ by (iii). Hence every element of U has an additive inverse in U.

The other parts of the definition of a vector space, such as associativity and commutativity, are automatically satisfied for U, because they hold on the larger space V. Thus U is a vector space and hence is a subspace of V. □

The next result states that a subspace of a subspace is a subspace.

Lemma 3.7. *If $U \leq V$ and $W \leq U$, then $W \leq V$.*

Proof. This is immediate from the definition of a subspace. □

The next result shows that the intersection of subspaces is a subspace.

Lemma 3.8. *Let $\{U_i \mid i \in I\}$ be a collection of subspaces of V. Then $\bigcap_{i \in I} U_i \leq V$.*

Proof. Let $U = \bigcap_{i \in I} U_i$.

(i) Since each $U_i \leq V$, $\mathbf{0} \in U_i$ so $\mathbf{0} \in U$.

(ii) Let $u, w \in U$, then $u, w \in U_i$. Since $U_i \leq V$, $u + w \in U_i$ so $u + w \in U$.

(iii) Let $\lambda \in \mathbf{F}$, $u \in U$, then $u \in U_i$. Since $U_i \leq V$, $\lambda u \in U_i$ so $\lambda u \in U$.

□

Definition 3.9 (Sum of subsets)**.** Suppose $U_1, \ldots, U_n \subset V$. The ***sum*** of $U_1, \ldots, U_n$ is the set of all possible sums of elements of $U_1, \ldots, U_n$:

$$U_1 + \cdots + U_n := \{u_1 + \cdots + u_n \mid u_i \in U_i\}.$$

Example.

- Let $U = \{(x, 0, 0) \in \mathbf{F}^3 \mid x \in F\}$ and $W = \{(0, y, 0) \in \mathbf{F}^3 \mid y \in \mathbf{F}\}$. Then

$$U + W = \{(x, y, 0) \mid x, y \in \mathbf{F}\}.$$

- Let $U = \{(x, x, y, y) \in \mathbf{F}^4 \mid x, y \in \mathbf{F}\}$ and $W = \{(x, x, x, y) \in \mathbf{F}^4 \mid x, y \in \mathbf{F}\}$. Then

$$U + W = \{(x, x, y, z) \in \mathbf{F}^4 \mid x, y, z \in \mathbf{F}\}.$$

The next result states that the sum of subspaces is a subspace, and is in fact the smallest subspace containing all the summands.

Proposition 3.10. *Suppose $U_1, \ldots, U_n \leq V$. Then $U_1 + \cdots + U_n$ is the smallest subspace of V containing $U_1, \ldots, U_n$.*

Proof. It is easy to see that $\mathbf{0} \in U_1 + \cdots + U_n$ and that $U_1 + \cdots + U_n$ is closed under addition and scalar multiplication. Hence by the subspace test, $U_1 + \cdots + U_n \leq V$.

Let M be the smallest subspace of V containing $U_1, \ldots, U_n$. We want to show that $U_1 + \cdots + U_n = M$. To do so, we show double inclusion.

$\boxed{\supset}$ For all $u_i \in U_i$ $(1 \leq i \leq n)$,

$$u_i = \mathbf{0} + \cdots + \mathbf{0} + u_i + \mathbf{0} + \cdots + \mathbf{0} \in U_1 + \cdots + U_n,$$

where all except one of the u's are $\mathbf{0}$. Thus $U_i \subset U_1 + \cdots + U_n$ for $1 \leq i \leq n$. Hence $M \subset U_1 + \cdots + U_n$.

$\boxed{\subset}$ Conversely, every subspace of V containing $U_1, \ldots, U_n$ contains $U_1 + \cdots + U_n$ (because subspaces must contain all finite sums of their elements). Hence $U_1 + \cdots + U_n \subset M$. □

Definition 3.11 (Direct sum). Suppose $U_1, \dots, U_n \leq V$. We say $U_1 + \cdots + U_n$ is a ***direct sum*** if each element of $U_1 + \cdots + U_n$ can be written in only one way as a sum $u_1 + \cdots + u_n$, $u_i \in U_i$. In this case, we denote the sum as

$$U_1 \oplus \cdots \oplus U_n.$$

Remark. When required to show that $V = \bigoplus_{i=1}^{n} U_i$, we have to show (i) $V = \sum_{i=1}^{n} U_i$ and (ii) $\sum_{i=1}^{n} U_i$ is a direct sum.

This is usually achieved by showing that every $v \in V$ can be uniquely expressed as $\sum_{i=1}^{n} u_i$ for some $u_i \in U_i$.

Example.

- Suppose $U = \{(x, y, 0) \in \mathbf{F}^3 \mid x, y \in \mathbf{F}\}$ and $W = \{(0, 0, z) \in \mathbf{F}^3 \mid z \in \mathbf{F}\}$. Then $\mathbf{F}^3 = U \oplus W$.
- Suppose U_i is the subspace of $\mathbf{F}^n$ of those vectors whose coordinates are all 0 except for the i-th coordinate; that is, $U_i = \{(0, \dots, 0, x, 0, \dots, 0) \in \mathbf{F}^n \mid x \in \mathbf{F}\}$. Then $\mathbf{F}^n = U_1 \oplus \cdots \oplus U_n$.

The definition of direct sum requires every vector in the sum to have a unique representation as an appropriate sum. The next result shows that when deciding whether a sum of subspaces is a direct sum, we only need to consider whether $\mathbf{0}$ can be uniquely written as an appropriate sum.

Lemma 3.12 (Condition for direct sum). *Suppose $V_1, \dots, V_n \leq V$. Then $V_1 \oplus \cdots \oplus V_n$ if and only if $v_1 + \cdots + v_n = \mathbf{0}$ implies $v_1 = \cdots = v_n = \mathbf{0}$.*

Proof.

$\boxed{\text{(i)} \implies \text{(ii)}}$ Suppose $V_1 + \cdots + V_n$ is a direct sum. Then by the definition of direct sum, the only way to write $\mathbf{0}$ as a sum $u_1 + \cdots + u_n$ is by taking $u_i = \mathbf{0}$.

$\boxed{\text{(ii)} \implies \text{(i)}}$ Suppose that the only way to write $\mathbf{0}$ as a sum $v_1 + \cdots + v_n$ by taking $v_1 = \cdots = v_n = \mathbf{0}$.

To show that $v \in V_1 + \cdots + V_n$ is a direct sum, let $v \in V_1 + \cdots + V_n$. Then

$$v = v_1 + \cdots + v_n \tag{I}$$

for some $v_i \in V_i$. To show that this representation is unique, suppose

$$v = v_1' + \cdots + v_n' \tag{II}$$

for some $v_i' \in V_i$. Substracting (II) from (I) gives

$$\mathbf{0} = (v_1 - v_1') + \cdots + (v_n - v_n').$$

Since $v_i - v_i' \in V_i$, the equation above implies $v_i - v_i' = \mathbf{0}$, so $v_i = v_i'$. Hence there is only one unique way to represent $v_1 + \cdots + v_n$. □

The next result provides a characterisation for direct sum.

Lemma 3.13. *Suppose $U, W \leq V$. Then $U + W$ is a direct sum if and only if $U \cap W = \{\mathbf{0}\}$.*

Proof.

$\boxed{\Longrightarrow}$ Suppose that $U + W$ is a direct sum. Let $v \in U \cap W$, we will show that $v = \mathbf{0}$.

Note that $\mathbf{0} = v + (-v)$, where $v \in U$, $-v \in W$. By the unique representation of $\mathbf{0}$ as the sum of a vector in U and a vector in W, we must have $v = \mathbf{0}$. Hence $U \cap W = \{\mathbf{0}\}$.

$\boxed{\Longleftarrow}$ Suppose $U \cap W = \{\mathbf{0}\}$. Suppose $u \in U$, $w \in W$, and $0 = u + w$. $u = -w \in W$, thus $u \in U \cap W$, so $u = w = \mathbf{0}$. By 3.12, $U + W$ is a direct sum. □

3.3 Span and Linear Independence

Definition 3.14 (Linear combination). We say $v \in V$ is a ***linear combination*** of $v_1, \ldots, v_n \in V$ if there exists $a_1, \ldots, a_n \in \mathbf{F}$ such that

$$\begin{aligned} v &= a_1 v_1 + \cdots + a_n v_n \\ &= \sum_{i=1}^{n} a_i v_i. \end{aligned}$$

Definition 3.15 (Span). The ***span*** of $\{v_1, \ldots, v_n\}$ is the set of all linear combinations of $v_1, \ldots, v_n$:

$$\operatorname{span}(v_1, \ldots, v_n) := \{a_1 v_1 + \cdots + a_n v_n \mid a_i \in \mathbf{F}\}.$$

We say $v_1, \ldots, v_n$ *spans* V if $\operatorname{span}(v_1, \ldots, v_n) = V$.
If $S \subset V$ is such that $\operatorname{span}(S) = V$, we say S *spans* V, and S is a *spanning set* for V:

$$\operatorname{span}(S) := \{a_1 v_1 + \cdots + a_n v_n \mid v_i \in S, a_i \in \mathbf{F}\}.$$

By convention, we define $\operatorname{span}(\emptyset) = \{\mathbf{0}\}$.

Proposition 3.16. $\operatorname{span}(v_1, \ldots, v_n)$ *in* V *is the smallest subspace of* V *containing* $v_1, \ldots, v_n$.

Proof. First we show that $\operatorname{span}(v_1, \ldots, v_n) \leq V$, using the subspace test.

(i) $\mathbf{0} = 0v_1 + \cdots + 0v_n \in \operatorname{span}(v_1, \ldots, v_n)$

(ii) $(a_1 v_1 + \cdots + a_n v_n) + (c_1 v_1 + \cdots + c_n v_n) = (a_1 + c_1)v_1 + \cdots + (a_n + c_n)v_n \in \operatorname{span}(v_1, \ldots, v_n)$, so $\operatorname{span}(v_1, \ldots, v_n)$ is closed under addition.

(iii) $\lambda(a_1 v_1 + a_n v_n) = (\lambda a_1)v_1 + \cdots + (\lambda a_n)v_n \in \operatorname{span}(v_1, \ldots, v_n)$, so $\operatorname{span}(v_1, \ldots, v_n)$ is closed under scalar multiplication.

Let M be the smallest vector subspace of V containing $v_1, \ldots, v_n$. We claim that $M = \operatorname{span}(v_1, \ldots, v_n)$.

$\boxed{\subset}$ Each v_i is a linear combination of $v_1, \ldots, v_n$, as

$$v_i = 0 \cdot v_1 + \cdots + 0 \cdot v_{i-1} + 1 \cdot v_i + 0 \cdot v_{i+1} + \cdots + 0 \cdot v_n,$$

so by the definition of the span as the collection of all linear combinations of $v_1, \ldots, v_n$, we have that $v_i \in \operatorname{span}(v_1, \ldots, v_n)$. But M is the smallest vector subspace containing $v_1, \ldots, v_n$, so

$$M \subset \operatorname{span}(v_1, \ldots, v_n).$$

$\boxed{\supset}$ Since $v_i \in M$ $(1 \leq i \leq n)$ and M is a vector subspace (closed under addition and scalar multiplication), it follows that

$$a_1 v_1 + \cdots + a_n v_n \in M$$

for all $a_i \in \mathbf{F}$ (i.e., M contains all linear combinations of $v_1, \ldots, v_n$). Thus

$$\operatorname{span}(v_1, \ldots, v_n) \subset M.$$

□

Definition 3.17 (Finite-dimensional)**.** We say V is ***finite-dimensional*** if it has a finite spanning set; otherwise, it is *infinite-dimensional*.

Example. For positive integer n, $\mathbf{F}^n$ is finite-dimensional.

Proof. Suppose $(x_1, x_2, \ldots, x_n) \in \mathbf{F}^n$, then

$$(x_1, x_2, \ldots, x_n) = x_1(1, 0, \ldots, 0) + x_2(0, 1, \ldots, 0) + \cdots + x_n(0, 0, \ldots, 1)$$

so

$$(x_1, \ldots, x_n) \in \operatorname{span}\left((1, 0, \ldots, 0), (0, 1, \ldots, 0), \ldots, (0, \ldots, 0, 1)\right).$$

The vectors $(1, 0, \ldots, 0), (0, 1, \ldots, 0), \ldots, (0, \ldots, 0, 1)$ spans $\mathbf{F}^n$, so $\mathbf{F}^n$ is finite-dimensional. □

Definition 3.18 (Linear independence)**.** We say $v_1, \ldots, v_n$ are ***linearly independent*** in V if

$$a_1 v_1 + \cdots + a_n v_n = \mathbf{0} \implies a_1 = \cdots = a_n = 0.$$

Otherwise, the vectors are *linearly dependent*.
We say $S \subset V$ is linearly independent if every finite subset of S is linearly independent.

Lemma 3.19 (Compare coefficients)**.** *Let $v_1, \ldots, v_n$ be linearly independent in V. Then*

$$a_1 v_1 + \cdots + a_n v_n = b_1 v_1 + \cdots + b_n v_n$$

if and only if $a_i = b_i$ $(1 \leq i \leq n)$.

Proof. Exercise. □

The following are easy consequences of the definition.

1. Any set which contains a linearly dependent set is linearly dependent.
2. Any subset of a linearly independent set is linearly independent.

3. Any set which contains $\mathbf{0}$ is linearly dependent.

4. A set S of vectors is linearly independent if and only if each finite subset of S is linearly independent.

The following result will often be useful; (i) states that given a linearly dependent set of vectors, one of the vectors is in the span of the previous ones; furthermore (ii) states that we can throw out that vector without changing the span of the original set.

Lemma 3.20 (Linear dependence lemma). *Suppose $v_1, \ldots, v_n$ are linearly dependent in V. Then there exists v_i such that the following hold:*

(i) $v_i \in \operatorname{span}(v_1, \ldots, v_{i-1})$

(ii) $\operatorname{span}(v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n) = \operatorname{span}(v_1, \ldots, v_n)$

Proof.

(i) Since $v_1, \ldots, v_n$ are linearly dependent, there exists $a_1, \ldots, a_n \in \mathbf{F}$, not all 0, such that

$$a_1 v_1 + \cdots + a_n v_n = 0.$$

Take $i = \max\{1, \ldots, n\}$ such that $a_i \neq 0$. Then

$$v_i = -\frac{a_1}{a_i} v_1 - \cdots - \frac{a_{i-1}}{a_i} v_{i-1},$$

since $a_{i-1}, \ldots, a_n = 0$. Thus v_i can be written as a linear combination of $v_1, \ldots, v_{i-1}$, so $v_i \in \operatorname{span}(v_1, \ldots, v_{i-1})$.

(ii) Now suppose i is such that $v_i \in \operatorname{span}(v_1, \ldots, v_{i-1})$. Then there exists $b_1, \ldots, b_{i-1} \in \mathbf{F}$ be such that

$$v_i = b_1 v_1 + \cdots + b_{i-1} v_{i-1}. \tag{I}$$

Suppose $u \in \operatorname{span}(v_1, \ldots, v_n)$. Then there exists $c_1, \ldots, c_n \in \mathbf{F}$ such that

$$u = c_1 v_1 + \cdots + c_n v_n. \tag{II}$$

Substituting (I) into (II) gives

$$\begin{aligned} u &= c_1 v_1 + \cdots + c_{i-1} v_{i-1} + c_i v_i + c_{i+1} v_{i+1} + \cdots + c_n v_n \\ &= c_1 v_1 + \cdots + c_{i-1} v_{i-1} + c_i (b_1 v_1 + \cdots + b_{i-1} v_{i-1}) + c_{i+1} v_{i+1} + \cdots + c_n v_n \\ &= c_1 v_1 + \cdots + c_{i-1} v_{i-1} + c_i b_1 v_1 + \cdots + c_i b_{i-1} v_{i-1} + c_{i+1} v_{i+1} + \cdots + c_n v_n \\ &= (c_1 + b c_i) v_1 + \cdots + (c_{i-1} + b_{i-1} c_i) v_{i-1} + c_{i+1} v_{i+1} + \cdots + c_n v_n. \end{aligned}$$

Thus $u \in \text{span}(v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n)$. This shows that removing v_i from $v_1, \ldots, v_n$ does not change the span of the set.

□

The next result says that no linearly independent set in V is longer than a spanning set in V.

Proposition 3.21. *In a finite-dimensional vector space, the length of every linearly independent set of vectors is less than or equal to the length of every spanning set of vectors.*

That is,

$$\{\text{linearly independent}\} \leq \{\text{spanning set}\}. \tag{3.1}$$

Proof. Suppose $A = \{u_1, \ldots, u_m\}$ is linearly independent in V, $B = \{w_1, \ldots, w_n\}$ spans V. We want to show that $m \leq n$.

We do so through the process described below with m steps; in each step, we add one of the u's and remove one of the w's.

Step 1 Since B spans V, if we add any other vector to B, we will get a linearly dependent set, since this newly added vector can, by the definition of a span, be expressed as a linear combination of the vectors in B. In particular, if we add u_1 to B, then the new set

$$\{u_1, w_1, \ldots, w_n\}$$

is linearly dependent.

By the linear independence lemma, one of the vectors in the above set is a linear combination of the previous vectors. Since $\{u_1, \ldots, u_m\}$ is linearly independent, $u_1 \neq \mathbf{0}$ so $u_1 \notin \text{span}\{\,\} = \{\mathbf{0}\}$. Hence the linear dependence lemma implies we can remove one of the w's, so that the new set B (of length n) consisting of u_1 and the remaining w's spans V.

Step i $(2 \leq i \leq m)$ The set B (of length n) from step $i-1$ spans V. In particular, u_i is in the span of B. If we add u_i to B, placing it just after $u_1, \ldots, u_{i-1}$, then the new set (of length $n+1$)

$$\{u_1, \ldots, u_{i-1}, u_i, w\text{'s}\}$$

is linearly dependent.

By the linear dependence lemma, one of the vectors in this set is in the span of the previous ones. Since $u_1, \ldots, u_i$ are linearly independent, this vector cannot be one of the u's. Hence there still must be at least one remaining w at this step. We can remove from our new set (after adjoining u_i in the proper place) a w that is a linear combination of the previous vectors in the set, so that the new set B (of length n) consisting of $u_1, \ldots, u_i$ and the remaining w's spans V.

After step m, we have added *all* the u's and the process stops. At each step as we add a u to B, the linear dependence lemma implies that there is some w to remove. Hence there must be at least as many w's as u's. □

We can use this result to show, without any computations, that certain sets are not linearly independent and that certain sets do not span a given vector space.

Example.

- $(1,0,0),(0,1,0),(0,0,1)$ spans $\mathbb{R}^3$. Thus no set of length larger than three is linearly independent in $\mathbb{R}^3$.
- $(1,0,0,0),(0,1,0,0),(0,0,1,0),(0,0,0,1)$ is linearly independent in $\mathbb{R}^4$. Thus no set of length less than four spans $\mathbb{R}^4$.

Our intuition suggests that every subspace of a finite-dimensional vector space should also be finite-dimensional. We now prove that this intuition is correct.

Proposition 3.22. *Every subspace of a finite-dimensional vector space is finite-dimensional.*

Proof. Suppose V is finite-dimensional, $U \leq V$. To show that U is finite-dimensional, we shall construct a spanning set of vectors in U, via the following steps.

Step 1 If $U = \{\mathbf{0}\}$, then U is finite-dimensional and we are done.

Otherwise, choose $v_1 \in U$, $v_1 \neq \mathbf{0}$ and add it to our set of vectors.

Step i Our set so far is $\{v_1, \ldots, v_{i-1}\}$.

If $U = \text{span}(v_1, \ldots, v_{i-1})$, then U is finite-dimensional and we are done.

Otherwise, choose $v_i \in U$ such that $v_i \notin \text{span}(v_1, \ldots, v_{i-1})$ and add it to our set.

After each step, we have constructed a set of vectors such that no vector in this set is in the span of the previous vectors; by the linear dependence lemma, our constructed set is linearly independent.

By 3.21, this linearly independent set cannot be longer than any spanning set of V. Thus the process must terminate after a finite number of steps, and we have constructed a spanning set of U. Hence U is finite-dimensional. □

3.4 Bases

Definition 3.23 (Basis). We say $B = \{v_1 \ldots, v_n\}$ is a ***basis*** of V if

(i) B is linearly independent in V, and

(ii) B is a spanning set of V.

Example (Standard basis). Let $\mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)$, where the i-th coordinate is 1. Then $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ is a basis of $\mathbf{F}^n$, known as the *standard basis* of $\mathbf{F}^n$.

The next result helps explain why bases are useful.

Lemma 3.24 (Criterion for basis). *Let $B = \{v_1, \ldots, v_n\}$ be a set of vectors in V. Then B is a basis of V if and only if every $v \in V$ can be uniquely expressed as a linear combination of $v_1, \ldots, v_n$.*

Proof.

$\boxed{\Longrightarrow}$ Let $v \in V$. Since B is a basis of V, there exist $a_1, \ldots, a_n \in \mathbf{F}$ such that

$$v = a_1 v_1 + \cdots + a_n v_n. \tag{I}$$

To show that the representation is unique, suppose that $c_1, \ldots, c_n \in \mathbf{F}$ also satisfy

$$v = c_1 v_1 + \cdots + c_n v_n. \tag{II}$$

Subtracting (II) from (I) gives

$$\mathbf{0} = (a_1 - c_1)v_1 + \cdots + (a_n - c_n)v_n.$$

Since $v_1, \ldots, v_n$ are linearly independent, we have $a_i - c_i = 0$, or $a_i = c_i$ for all i.

$\boxed{\Longleftarrow}$ Suppose every $v \in V$ can be uniquely expressed as a linear combination of $v_1, \ldots, v_n$. This implies that B spans V.

To show that B is linearly independent, suppose that $a_1, \ldots, a_n \in \mathbf{F}$ are such that

$$a_1 v_1 + \cdots + a_n v_n = \mathbf{0}.$$

Since $\mathbf{0}$ can be uniquely expressed as a linear combination of $v_1, \ldots, v_n$, we have $a_1 = \cdots = a_n = 0$, thus B is linearly independent.

Since B is linearly independent and spans V, we conclude that B is a basis of V. □

A spanning set in a vector space may not be a basis because it is not linearly independent. The next result says that given any spanning set, we can *remove* some vectors so that the remaining set is linearly independent and still spans the vector space.

Lemma 3.25. *In a vector space, every spanning set can be reduced to a basis.*

Proof. Suppose $B = \{v_1, \ldots, v_n\}$ spans V. We want to remove some vectors from B so that the remaining vectors form a basis of V. We do this through the multistep process described below.

Step 1 If $v_1 = \mathbf{0}$, delete v_1 from B. If $v_1 \neq \mathbf{0}$, leave B unchanged.

Step i $(2 \leq i \leq n)$ If $v_i \in \text{span}(v_1, \ldots, v_{i-1})$, delete v_i from B. If $v_i \notin \text{span}(v_1, \ldots, v_{i-1})$, leave B unchanged.

Since we only delete vectors from B that are in the span of the previous vectors, by the linear dependence lemma, the set B still spans V.

The process ensures that no vector in B is in the span of the previous ones. By the linear dependence lemma, B is linearly independent.

Since B is linearly independent and spans V, we conclude that B is a basis of V. □

Corollary 3.26. *Every finite-dimensional vector space has a basis.*

Proof. We prove by construction. Suppose V is finite-dimensional. By definition, there exists a spanning set of vectors in V. By 3.25, the spanning set can be reduced to a basis. □

We now show that given any linearly independent set, we can *add* some vectors so that the extended set is still linearly independent but also spans the space.

Lemma 3.27. *In a finite-dimensional vector space, every linearly independent set can be extended to a basis.*

Proof. Suppose $\{u_1, \ldots, u_m\}$ is linearly independent in V, and $\{w_1, \ldots, w_n\}$ spans V. Then the set

$$\{u_1, \ldots, u_m, w_1, \ldots, w_n\}$$

spans V. By 3.25, we can reduce this set to a basis of V consisting $u_1, \ldots, u_m$ (since $u_1, \ldots, u_m$ are linearly independent, $u_i \notin \text{span}(u_1, \ldots, u_{i-1})$ for all i, so none of the u_i's are deleted in the process), and some of the w_i's. □

We now show that every subspace of a finite-dimensional vector space can be paired with another subspace to form a direct sum of the whole space.

Corollary 3.28. *Suppose V is finite-dimensional, $U \leq V$. Then there exists $W \leq V$ such that $V = U \oplus W$.*

Proof. Since V is finite-dimensional and $U \leq V$, by 3.22, U is finite-dimensional. By 3.26, U has a basis, say $B = \{u_1, \ldots, u_n\}$.

Since B is linearly independent, by 3.27, B can be extended to a basis of V, say

$$\{u_1, \ldots, u_n, w_1, \ldots, w_n\}.$$

Claim. $W = \operatorname{span}(w_1, \ldots, w_n)$.

We need to show that $V = U \oplus W$; by 3.12, we need to show (i) $V = U + W$, and (ii) $U \cap W = \{\mathbf{0}\}$.

(i) Let $v \in V$. Since $\{u_1, \ldots, u_n, w_1, \ldots, w_n\}$ spans V, there exists $a_1, \ldots, a_n, b_1, \ldots, b_n \in \mathbf{F}$ such that

$$v = a_1u_1 + \cdots + a_nu_n + b_1w_1 + \cdots + b_nw_n.$$

Take $u = a_1u_1 + \cdots + a_nu_n \in U$, $w = b_1w_1 + \cdots + b_nw_n \in W$. Then $v = u + w \in U + W$, so $V = U + W$.

(ii) Let $v \in U \cap W$. Since $v \in U$, v can be written as a linear combination of $u_1, \ldots, u_n$:

$$v = a_1u_1 + \cdots + a_nu_n. \tag{I}$$

Since $v \in W$, v can be written as a linear combination of $w_1, \ldots, w_n$:

$$v = b_1w_1 + \cdots + b_nw_n. \tag{II}$$

Subtracting (II) from (I) gives

$$\mathbf{0} = a_1u_1 + \cdots + a_nu_n - b_1w_1 - \cdots - b_nw_n.$$

Since $u_1, \ldots, u_n, w_1, \ldots, w_n$ are linearly independent, we have $a_i = b_i = 0$ for all i. Thus $v = \mathbf{0}$, so $U \cap W = \{\mathbf{0}\}$.

□

3.5 Dimension

Lemma 3.29. *Any two bases of a finite-dimensional vector space have the same length.*

Proof. Suppose V is finite-dimensional, let B_1 and B_2 be two bases of V. By definition, B_1 is linearly independent in V, and B_2 spans V, so by 3.21, $|B_1| \leq |B_2|$.

Similarly, by definition, B_2 is linearly independent in V and B_1 spans V, so $|B_2| \leq |B_1|$.

Since $|B_1| \leq |B_2|$ and $|B_2| \leq |B_1|$, we have $|B_1| = |B_2|$, as desired. □

Since any two bases of a finite-dimensional vector space have the same length, we can formally define the dimension of such spaces.

Definition 3.30 (Dimension). The ***dimension*** of V is the length of any basis of V, denoted by $\dim V$.

We define $\dim\{\mathbf{0}\} = 0$.

Lemma 3.31 (Dimension of subspace). *Suppose V is finite-dimensional, $U \leq V$. Then* $\dim U \leq \dim V$.

Proof. Since V is finite-dimensional and $U \leq V$, U is finite-dimensional. Let B_U be a basis of U, and B_V be a basis of V.

By definition, B_U is linearly independent in V, and B_V spans V. By 3.21, $|B_U| \leq |B_V|$, so

$$\dim U = |B_U| \leq |B_V| = \dim V.$$

□

To check that a set of vectors is a basis, we must show that it is linearly independent and that it spans the vector space. The next result shows that if the set has the *right length*, then we only need to check that it satisfies one of the two required properties.

Proposition 3.32. *Suppose V is finite-dimensional. Then*

(i) every linearly independent set of vectors in V with length $\dim V$ is a basis of V;

(ii) every spanning set of vectors in V with length $\dim V$ is a basis of V.

Proof.

(i) Suppose $\dim V = n$, and $\{v_1, \ldots, v_n\}$ is linearly independent in V.

By 3.27, $\{v_1,\dots,v_n\}$ can be extended to a basis of V. However, every basis of V has length n, which means no elements are added to $\{v_1,\dots,v_n\}$. Hence $\{v_1,\dots,v_n\}$ is a basis of V.

(ii) Suppose $\dim V = n$, and $\{v_1,\dots,v_n\}$ spans V.

By 3.25, $\{v_1,\dots,v_n\}$ can be reduced to a basis of V. However, every basis of V has length n, which means no elements are removed from $\{v_1,\dots,v_n\}$. Hence $\{v_1,\dots,v_n\}$ is a basis of V.

□

Corollary 3.33. *Suppose V is finite-dimensional, $U \leq V$. If $\dim U = \dim V$, then $U = V$.*

Proof. Let $\dim U = \dim V = n$, let $\{u_1,\dots,u_n\}$ be a basis of U.

Then $\{u_1,\dots,u_n\}$ is linearly independent in V (because it is a basis of U) of length $\dim V$. By 3.32, $\{u_1,\dots,u_n\}$ is a basis of V. In particular every vector in V is a linear combination of $u_1,\dots,u_n$. Thus $U = V$. □

The next result gives a formula for the dimension of the sum of two subspaces of a finite-dimensional vector space.

Lemma 3.34 (Dimension of sum). *Suppose V is finite-dimensional, $U_1, U_2 \leq V$. Then*

$$\dim(U_1 + U_2) = \dim U_1 + \dim U_2 - \dim(U_1 \cap U_2). \tag{3.2}$$

Proof. Let $\{u_1,\dots,u_m\}$ be a basis of $U_1 \cap U_2$; thus $\dim(U_1 \cap U_2) = m$.

Since $\{u_1,\dots,u_m\}$ is a basis of $U_1 \cap U_2$, it is linearly independent in U_1. By 3.27, $\{u_1,\dots,u_m\}$ can be extended to a basis $\{u_1,\dots,u_m,v_1,\dots,v_j\}$ of U_1; thus $\dim U_1 = m + j$. Similarly, extend $\{u_1,\dots,u_m\}$ to a basis $\{u_1,\dots,u_m,v_1,\dots,v_k\}$ of U_2; thus $\dim U_2 = m + k$.

We will show that

$$\{u_1,\dots,u_m,v_1,\dots,v_j,w_1,\dots,w_k\}$$

is a basis of $U_1 + U_2$. This will complete the proof because then we will have

$$\begin{aligned}\dim(U_1 + U_2) &= m + j + k \\ &= (m + j) + (m + k) - m \\ &= \dim U_1 + \dim U_2 - \dim(U_1 \cap U_2).\end{aligned}$$

We just need to show that $\{u_1,\dots,u_m,v_1,\dots,v_j,w_1,\dots,w_k\}$ is linearly independent. To prove this, suppose

$$a_1u_1 + \cdots + a_mu_m + b_1v_1 + \cdots + b_jv_j + c_1w_1 + \cdots + c_kw_k = \mathbf{0}, \tag{I}$$

where $a_i, b_i, c_i \in \mathbf{F}$. We need to show that $a_i = b_i = c_i = 0$ for all i. (I) can be rewritten as

$$c_1w_1 + \cdots + c_kw_k = -a_1u_1 - \cdots - a_mu_m - b_1v_1 - \cdots - b_jv_j,$$

which shows that $c_1w_1 + \cdots + c_kw_k \in U_1$. But actually all the w_i's are in U_2, so $c_1w_1 + \cdots + c_kw_k \in U_2$. Thus $c_1w_1 + \cdots + c_kw_k \in U_1 \cap U_2$. Then we can write

$$c_1w_1 + \cdots + c_kw_k = d_1u_1 + \cdots + d_mu_m$$

for some $d_i \in \mathbf{F}$. But $u_1, \ldots, u_m, w_1, \ldots, w_k$ are linearly independent, so $c_i = d_i = 0$ for all i. Thus our original equation (I) becomes

$$a_1u_1 + \cdots + a_mu_m + b_1v_1 + \cdots + b_jv_j = \mathbf{0}.$$

Since $u_1, \ldots, u_m, v_1, \ldots, v_j$ are linearly independent, $a_i = b_i = 0$ for all i, as desired. □

Exercises

Exercise 3.1 ([Axl24] 1C Q12)**.** Suppose W is a vector space over $\mathbf{F}$, V_1 and V_2 are subspaces of W. Show that $V_1 \cup V_2$ is a vector space over $\mathbf{F}$ if and only if $V_1 \subset V_2$ or $V_2 \subset V_1$.

Solution. The backward direction is trivial. We focus on proving the forward direction. Supppse otherwise, then $V_1 \setminus V_2 \neq \emptyset$ and $V_2 \setminus V_1 \neq \emptyset$. Pick $v_1 \in V_1 \setminus V_2$ and $v_2 \in V_2 \setminus V_1$. Then

$$\begin{aligned} v_1, v_2 \in V_1 \cup V_2 &\implies v_1 + v_2 \in V_1 \cup V_2 \\ &\implies v_2, v_1 + v_2 \in V_2 \\ &\implies v_1 = (v_1 + v_2) - v_2 \in V_2 \end{aligned}$$

which is a contradiction. □

Exercise 3.2 ([Axl24] 1C Q13)**.** Suppose W is a vector space over $\mathbf{F}$, V_1, V_2, V_3 are subspaces of W. Then $V_1 \cup V_2 \cup V_3$ is a vector space over $\mathbf{F}$ if and only if one of the V_i contains the other two.

Solution. We prove the forward direction. Suppose otherwise, then $v_1 \in V_1 \setminus (V_2 + V_3)$, $v_2 \in V_2 \setminus (V_1 + V_3)$, $v_3 \in V_3 \setminus (V_1 + V_2)$. Consider

$$\{v_1 + v_2 + v_3, v_1 + v_2 + 2v_3, v_1 + 2v_2 + v_3, v_1 + 2v_2 + 2v_3\} \subset V_1 \cup V_2 \cup V_3$$

Then

$$\begin{aligned} &(v_1 + v_2 + 2v_3) - (v_1 + v_2 + v_3) = v_3 \notin V_1 + V_2 \\ &\implies v_1 + v_2 + v_3 \notin V_1 + V_2 \quad \text{or} \quad v_1 + v_2 + 2v_3 \notin V_1 + V_2 \\ &\implies v_1 + v_2 + v_3 \in V_3 \quad \text{or} \quad v_1 + v_2 + 2v_3 \in V_3 \\ &\implies v_1 + v_2 \in V_3 \end{aligned}$$

Similarly,

$$\begin{aligned} &(v_1 + 2v_2 + 2v_3) - (v_1 + 2v_2 + v_3) = v_3 \notin V_1 + V_2 \\ &\implies v_1 + 2v_2 + v_3 \notin V_1 + V_2 \quad \text{or} \quad v_1 + 2v_2 + 2v_3 \notin V_1 + V_2 \\ &\implies v_1 + 2v_2 + v_3 \in V_3 \quad \text{or} \quad v_1 + 2v_2 + 2v_3 \in V_3 \\ &\implies v_1 + 2v_2 \in V_3 \end{aligned}$$

This implies $(v_1 + 2v_2) - (v_1 + v_2) = v_2 \in V_3$, a contradiction. □

Exercise 3.3 ([Axl24] 2A Q12)**.** Suppose $\{v_1,\dots,v_n\}$ is linearly independent in V, $w \in V$. Prove that if $\{v_1+w,\dots,v_n+w\}$ is linearly dependent, then $w \in \operatorname{span}(v_1,\dots,v_n)$.

Solution. If $\{v_1+w,\dots,v_n+w\}$ is linearly dependent, then there exists $a_1,\dots,a_n \in \mathbf{F}$, not all zero, such that

$$a_1(v_1+w)+\cdots+a_n(v_n+w)=0,$$

or

$$a_1v_1+\cdots+a_nv_n=-(a_1+\cdots+a_n)w.$$

Suppose otherwise, that $a_1+\cdots+a_n=0$. Then

$$a_1v_1+\cdots+a_nv_n=\mathbf{0},$$

but the linear independence of $\{v_1,\dots,v_n\}$ implies that $a_1=\cdots=a_n=0$, which is a contradiction. Hence we must have $a_1+\cdots+a_n\neq 0$, so we can write

$$w=-\frac{a_1}{a_1+\cdots+a_n}v_1-\cdots-\frac{a_n}{a_1+\cdots+a_n}v_n,$$

which is a linear combination of $v_1,\dots,v_n$. Thus by definition of span, $w \in \operatorname{span}(v_1,\dots,v_n)$. □

Exercise 3.4 ([Axl24] 2A Q14)**.** Suppose $\{v_1,\dots,v_n\} \subset V$. Let

$$w_i = v_1+\cdots+v_i \quad (i=1,\dots,n)$$

Show that $\{v_1,\dots,v_n\}$ is linearly independent if and only if $\{w_1,\dots,w_n\}$ is linearly independent.

Solution. Write

$$\begin{aligned} v_1 &= w_1 \\ v_2 &= w_2-w_1 \\ v_3 &= w_3-w_2 \\ &\vdots \\ v_n &= w_n-w_{n-1}. \end{aligned}$$

$\boxed{\Longrightarrow}$

$$a_1w_1+\cdots+a_nw_n=\mathbf{0}$$

for some $a_i \in \mathbf{F}$. Expressing w_i's as v_i's,

$$a_1v_1+a_2(v_1+v_2)+\cdots+a_n(v_1+\cdots+v_n)=0,$$

or

$$(a_1 + \cdots + a_n)v_1 + (a_2 + \cdots + a_n)v_2 + \cdots + a_n v_n = \mathbf{0}.$$

Since $v_1, \ldots, v_n$ are linearly independent,

$$\begin{aligned} a_1 + a_2 + \cdots + a_n &= 0 \\ a_2 + \cdots + a_n &= 0 \\ &\vdots \\ a_n &= 0 \end{aligned}$$

on solving simultaneously gives $a_1 = \cdots = a_n = 0$.

$\boxed{\Longleftarrow}$ Similar to the above. □

Exercise 3.5 ([Axl24] 2A Q18)**.** Prove that $\mathbf{F}^\infty$ is infinite-dimensional.

Solution. Suppose, for a contradiction, that $\mathbf{F}^\infty$ is finite-dimensional, i.e., there exists a finite spanning set $\{v_1, \ldots, v_n\}$. Let

$$\begin{aligned} e_1 &= (1, 0, \ldots) \\ e_2 &= (0, 1, 0, \ldots) \\ e_3 &= (0, 0, 1, 0, \ldots) \\ &\vdots \\ e_{n+1} &= (0, \ldots, 0, 1, 0, \ldots) \end{aligned}$$

where e_i has a 1 at the i-th coordinate, and 0's for the remaining coordinates. Let

$$a_1 e_1 + \cdots + a_{n+1} e_{n+1} = \mathbf{0}$$

for some $a_i \in \mathbf{F}$. Then

$$(a_1, a_2, \ldots, a_{n+1}, 0, 0, \ldots) = \mathbf{0}$$

so $a_1 = a_2 = \cdots = a_{n+1} = 0$. Thus $\{e_1, \ldots, e_{n+1}\}$ is a linearly independent set, of length $n+1$. However, $\{v_1, \ldots, v_n\}$ is a spanning set of length n. By 3.21, we have reached a contradiction. □

Exercise 3.6 ([Axl24] 2B Q5)**.** Suppose V is finite-dimensional, $U, W \leq V$ such that $V = U + W$. Prove that V has a basis in $U \cup W$.

Solution. Let $\{v_i\}_{i=1}^n$ denote the basis for V. By definition we have $v_i = u_i + w_i$ for some $u_i \in U$, $w_i \in W$. Then we have the spanning set of the vector space V $\sum_{i=1}^n a_i(u_i + w_i)$, which can be reduced to a basis by the lemma. □

Exercise 3.7 ([Axl24] 2B Q7)**.** Suppose $\{v_1, v_2, v_3, v_4\}$ is a basis of V. Prove that

$$\{v_1 + v_2, v_2 + v_3, v_3 + v_4, v_4\}$$

is also a basis of V.

Solution. We know that $\{v_1, v_2, v_3, v_4\}$ is linearly independent and spans V. Then there exist $a_i \in \mathbf{F}$ such that

$$a_1(v_1 + v_2) + a_2(v_2 + v_3) + a_3(v_3 + v_4) + a_4 v_4 = 0 \implies a_1 = a_2 = a_3 = a_4 = 0.$$

Write

$$\begin{aligned} &a_1(v_1 + v_2) + a_2(v_2 + v_3) + a_3(v_3 + v_4) + a_4 v_4 \\ &= a_1 v_1 + (a_1 + a_2)v_2 + (a_2 + a_3)v_3 + (a_3 + a_4)v_4, \end{aligned}$$

this shows the linear independence. To prove spanning, let $v \in V$, then

$$\begin{aligned} v &= a_1 v_1 + a_2 v_2 + a_3 v_3 + a_4 v_4 \\ &= a_1(v_1 + v_2) + (a_2 - a_1)(v_2 + v_3) + (a_3 - a_2)(v_3 + v_4) + (a_4 - a_3)v_4, \end{aligned}$$

which is a linear combination of $v_1 + v_2, v_2 + v_3, v_3 + v_4, v_4$. □

Exercise 3.8 ([Axl24] 2B Q10)**.** Suppose $U, W \leq V$ such that $V = U \oplus W$. Suppose also that $\{u_1, \ldots, u_m\}$ is a basis of U, $\{w_1, \ldots, w_n\}$ is a basis of W. Prove that

$$\{u_1, \ldots, u_m, w_1, \ldots, w_n\}$$

is a basis of V.

Solution. We know that this set is linearly independent (otherwise violating the direct sum assumption) so it suffices to prove the spanning. Let $v \in V$, then

$$v = u + w = \sum_{i=1}^{m} a_i u_i + \sum_{i=1}^{n} b_j w_j.$$

□

Exercise 3.9 ([Axl24] 2C Q8)**.**

Exercise 3.10 ([Axl24] 2C Q16)**.**

Exercise 3.11 ([Axl24] 2C Q17)**.** Suppose that $V_1, \dots, V_n \leq V$ are finite-dimensional. Prove that $V_1 + \cdots + V_n$ is finite-dimensional, and

$$\dim(V_1 + \cdots + V_n) \leq \dim V_1 + \cdots + \dim V_n.$$

Solution. We prove by induction on n. The base case is trivial. Assume the statement holds for k. Then for $k+1$, denoting $V_1 + \cdots + V_k = M_k$, we have that

$$\dim(M_k + V_{k+1}) \leq \dim V_1 + \cdots + \dim V_{k+1},$$

which is finite. □

4 Linear Maps

4.1 Vector Space of Linear Maps

Definition 4.1 (Linear map)**.** A ***linear map*** from V to W is a function $T : V \to W$ satisfying the following properties:

(i) $T(v+w) = Tv + Tw$ for all $v, w \in V$; (additivity)

(ii) $T(\lambda v) = \lambda T(v)$ for all $\lambda \in \mathbf{F}$, $v \in V$. (homogeneity)

Notation. If there is no ambiguity, we omit the parantheses and write Tv instead of $T(v)$.

Example. If V is any vector space, the *identity map* I, defined by $Iv = v$, is a linear map on V. The *zero map* 0, defined by $0v = v$, is a linear map on V.

Let $\mathscr{L}(V,W)$ denote the set of linear maps from V to W, and let $\mathscr{L}(V)$ denote the set of linear maps on V (from V to V).

Lemma. *$\mathscr{L}(V,W)$ is a vector space, with addition and scalar multiplication defined as follows: for all $S, T \in \mathscr{L}(V,W)$, $\lambda \in \mathbf{F}$,*

$$(S+T)(v) = Sv + Tv$$
$$(\lambda T)(v) = \lambda(Tv)$$

for all $v \in V$.

Definition 4.2 (Product of linear maps)**.** Suppose $T \in \mathscr{L}(U,V)$, $S \in \mathscr{L}(V,W)$. Define the ***product*** $ST \in \mathscr{L}(U,W)$ by

$$(ST)(u) := S(Tu) \quad (u \in U).$$

In other words, ST is just the usual composition $S \circ T$ of two functions.

Remark. ST is defined only when T maps into the domain of S.

Lemma 4.3 (Algebraic properties of products of linear maps)**.**

(i) Associativity: $(T_1T_2)T_3 = T_1(T_2T_3)$ *for all linear maps* T_1, T_2, T_3 *such that the products make sense (meaning that* T_3 *maps into the domain of* T_2*,* T_2 *maps into the domain of* T_1*)*

(ii) Identity: $TI = IT = T$ *for all* $T \in \mathscr{L}(V,W)$ *(the first* I *is the identity map on* V*, and the second* I *is the identity map on* W*)*

(iii) Distributive: $(S_1+S_2)T = S_1T + S_2T$ *and* $S(T_1+T_2) = ST_1 + ST_2$ *for all* $T, T_1, T_2 \in \mathscr{L}(U,V)$ *and* $S, S_1, S_2 \in \mathscr{L}(V,W)$

Proof. Exercise. □

Lemma 4.4. *Suppose* $T \in \mathscr{L}(V,W)$*. Then* $T(\mathbf{0}) = \mathbf{0}$*.*

Proof. By additivity,

$$T(\mathbf{0}) = T(\mathbf{0}+\mathbf{0}) = T(\mathbf{0}) + T(\mathbf{0}).$$

Add the additive inverse of $T(\mathbf{0})$ to each side of the equation to obtain $T(\mathbf{0}) = \mathbf{0}$. □

The existence part of the next result means that we can find a linear map that takes on whatever values we wish on the vectors in a basis. The uniqueness part of the next result means that a linear map is completely determined by its values on a basis.

Lemma 4.5 (Linear map lemma)**.** *Suppose* $\{v_1, \ldots, v_n\}$ *is a basis of* V*, and* $w_1, \ldots, w_n \in W$*. Then there exists a unique linear map* $T \in \mathscr{L}(V,W)$ *such that*

$$Tv_i = w_i \quad (i = 1, \ldots, n).$$

Proof.

Existence Define $T : V \to W$ as

$$T(c_1v_1 + \cdots + c_nv_n) = c_1w_1 + \cdots + c_nw_n,$$

for some $c_i \in \mathbf{F}$.

Since $\{v_1, \ldots, v_n\}$ is a basis of V, by 3.24, each $v \in V$ can be uniquely expressed as a linear combination of $v_1, \ldots, v_n$, thus the equation above does indeed define a function $T : V \to W$. For i $(1 \le i \le n)$, take $c_i = 1$ and the other c's equal to 0, then

$$T(0v_1 + \cdots + 1v_i + \cdots + 0v_n) = 0w_1 + \cdots + 1w_i + \cdots + 0w_n$$

which shows that $Tv_i = w_i$.

We now show that $T: V \to W$ is a linear map:

(i) For $u, v \in V$ with $u = a_1 v_1 + \cdots + a_n v_n$ and $c_1 v_1 + \cdots + c_n v_n$,

$$\begin{aligned} T(u+v) &= T\left((a_1+c_1)v_1 + \cdots + (a_n+c_n)v_n\right) \\ &= (a_1+c_1)w_1 + \cdots + (a_n+c_n)w_n \\ &= (a_1 w_1 + \cdots + a_n w_n) + (c_1 w_1 + \cdots + c_n w_n) \\ &= Tu + Tv. \end{aligned}$$

(ii) For $\lambda \in \mathbf{F}$ and $v = c_1 v_1 + \cdots + c_n v_n$,

$$\begin{aligned} T(\lambda v) &= T(\lambda c_1 v_1 + \cdots + \lambda c_n v_n) \\ &= \lambda c_1 w_1 + \cdots + \lambda c_n w_n \\ &= \lambda(c_1 w_1 + \cdots + c_n w_n) \\ &= \lambda T v. \end{aligned}$$

Uniqueness Suppose $T \in \mathscr{L}(V,W)$, and $Tv_i = w_i$ for $i = 1, \ldots, n$.

Let $c_i \in \mathbf{F}$. The homogeneity of T implies that $T(c_i v_i) = c_i w_i$. The additivity of T implies that

$$T(c_1 v_1 + \cdots + c_n v_n) = c_1 w_1 + \cdots + c_n w_n.$$

Thus T is uniquely determined on $\operatorname{span}\{v_1, \ldots, v_n\}$. Since $\{v_1, \ldots, v_n\}$ is a basis of V, this implies that T is uniquely determined on V. □

4.2 Kernel and Image

Definition 4.6 (Kernel)**.** Suppose $T \in \mathscr{L}(V,W)$. The ***kernel*** of T is

$$\ker T := \{v \in V \mid Tv = \mathbf{0}\}.$$

That is, $\ker T$ is the subset of V consisting of those vectors that T maps to $\mathbf{0}$.

Lemma. $\ker T \leq V$.

Proof.

(i) By 4.4, $T(\mathbf{0}) = \mathbf{0}$, so $\mathbf{0} \in \ker T$.

(ii) For all $v, w \in \ker T$, $T(v+w) = Tv + Tw = \mathbf{0} \implies v+w \in \ker T$, so $\ker T$ is closed under addition.

(iii) For all $v \in \ker T$, $\lambda \in \mathbf{F}$, $T(\lambda v) = \lambda Tv = \mathbf{0} \implies \lambda v \in \ker T$, so $\ker T$ is closed under scalar multiplication.

□

Definition 4.7 (Injectivity)**.** Suppose $T \in \mathscr{L}(V,W)$. We say T is ***injective*** if

$$Tu = Tv \implies u = v.$$

The next result provides a useful characterisation of injective linear maps.

Lemma 4.8. *Suppose $T \in \mathscr{L}(V,W)$. Then T is injective if and only if* $\ker T = \{\mathbf{0}\}$.

Proof.

$\boxed{\implies}$ Suppose T is injective. Let $v \in \ker T$, then

$$Tv = \mathbf{0} = T(\mathbf{0}) \implies v = \mathbf{0}$$

by the injectivity of T. Hence $\ker T = \{\mathbf{0}\}$ as desired.

$\boxed{\impliedby}$ Suppose $\ker T = \{\mathbf{0}\}$. Let $u, v \in V$ such that $Tu = Tv$. Then

$$T(u-v) = Tu - Tv = \mathbf{0}.$$

By definition of kernel, $u - v \in \ker T = \{\mathbf{0}\}$, so $u - v = \mathbf{0}$, which implies that $u = v$. Hence T is injective, as desired. □

Definition 4.9 (Image). Suppose $T \in \mathscr{L}(V,W)$. The ***image*** of T is

$$\operatorname{im} T := T(V) = \{Tv \mid v \in V\}.$$

That is, $\operatorname{im} T$ is the subset of W consisting of those vectors that are of the form Tv for some $v \in V$.

Lemma. $\operatorname{im} T \leq W$.

Proof.

(i) $T(\mathbf{0}) = \mathbf{0}$ implies that $\mathbf{0} \in \operatorname{im} T$.

(ii) For $w_1, w_2 \in \operatorname{im} T$, there exist $v_1, v_2 \in V$ such that $Tv_1 = w_1$ and $Tv_2 = w_2$. Then

$$w_1 + w_2 = Tv_1 + Tv_2 = T(v_1 + v_2) \in \operatorname{im} T \implies w_1 + w_2 \in \operatorname{im} T.$$

(iii) For $w \in \operatorname{im} T$ and $\lambda \in \mathbf{F}$, there exists $v \in V$ such that $Tv = w$. Then

$$\lambda w = \lambda Tv = T(\lambda v) \in \operatorname{im} T \implies \lambda w \in \operatorname{im} T.$$

□

Definition 4.10 (Surjectivity). Suppose $T \in \mathscr{L}(V,W)$. T is ***surjective*** if $\operatorname{im} T = W$.

Fundamental Theorem of Linear Maps

Theorem 4.11 (Fundamental theorem of linear maps). *Suppose V is finite-dimensional, $T \in \mathscr{L}(V,W)$. Then $\operatorname{im} T$ is finite-dimensional, and*

$$\dim V = \dim \ker T + \dim \operatorname{im} T. \tag{4.1}$$

If we define the *rank* of T as $\dim \operatorname{im} T$, and the *nullity* of T as $\dim \ker T$, then 4.11 is sometimes called the *rank-nullity theorem*.

Proof. Let $\{u_1, \ldots, u_m\}$ be basis of $\ker T$, then $\dim \ker T = m$. The linearly independent list $u_1, \ldots, u_m$ can be extended to a basis

$$\{u_1, \ldots, u_m, v_1, \ldots, v_n\}$$

of V, thus $\dim V = m + n$. To simultaneously show that $\operatorname{im} T$ is finite-dimensional and $\dim \operatorname{im} T = n$, we prove that $\{Tv_1, \ldots, Tv_n\}$ is a basis of $\operatorname{im} T$. Thus we need to show that the set (i) spans $\operatorname{im} T$, and (ii) is linearly independent.

(i) Let $v \in V$. Since $\{u_1, \ldots, u_m, v_1, \ldots, v_n\}$ spans V, we can write

$$v = a_1u_1 + \cdots + a_mu_m + b_1v_1 + \cdots + b_nv_n,$$

for some $a_i, b_i \in \mathbf{F}$. Applying T to both sides of the equation, and noting that $Tu_i = \mathbf{0}$ since $u_i \in \ker T$,

$$\begin{aligned} Tv &= T(a_1u_1 + \cdots + a_mu_m + b_1v_1 + \cdots + b_nv_n) \\ &= a_1\underbrace{Tu_1}_{\mathbf{0}} + \cdots + a_m\underbrace{Tu_m}_{\mathbf{0}} + b_1Tv_1 + \cdots + b_nv_n \\ &= b_1Tv_1 + \cdots + b_nTv_n \in \operatorname{im} T. \end{aligned}$$

Since every element of $\operatorname{im} T$ can be expressed as a linear combination of $Tv_1, \ldots, Tv_n$, we have that $\{Tv_1, \ldots, Tv_n\}$ spans $\operatorname{im} T$.

Moreover, since there exists a set of vectors that spans $\operatorname{im} T$, $\operatorname{im} T$ is finite-dimensional.

(ii) Suppose there exist $c_1, \ldots, c_n \in \mathbf{F}$ such that

$$c_1Tv_1 + \cdots + c_nTv_n = \mathbf{0}.$$

Then

$$T(c_1v_1 + \cdots + c_nv_n) = T(\mathbf{0}) = \mathbf{0},$$

which implies $c_1v_1 + \cdots + c_nv_n \in \ker T$. Since $\{u_1, \ldots, u_m\}$ is a spanning set of $\ker T$, we can write

$$c_1v_1 + \cdots + c_nv_n = d_1u_1 + \cdots + d_mu_m$$

for some $d_i \in \mathbf{F}$, or

$$c_1v_1 + \cdots + c_nv_n - d_1u_1 - \cdots - d_mu_m = \mathbf{0}.$$

Since $u_1, \ldots, u_m, v_1, \ldots, v_n$ are linearly independent, $c_i = d_i = 0$. Since $c_i = 0$, $\{Tv_1, \ldots, Tv_n\}$ is linearly independent.

□

No linear map from a finite-dimensional vector space to a "smaller" vector space can be injective:

Proposition 4.12. *Suppose V and W are finite-dimensional vector spaces, $\dim V > \dim W$. Then there does not exist $T \in \mathscr{L}(V, W)$ such that T is injective.*

Proof. Since W is finite-dimensional and $\operatorname{im} T \leq W$, by 3.31, we have that $\dim \operatorname{im} T \leq \dim W$.

Let $T \in \mathscr{L}(V,W)$. Then

$$\begin{aligned} \dim\ker T &= \dim V - \dim\operatorname{im} T && \text{[by fundamental theorem of linear maps]} \\ &\geq \dim V - \dim W > 0. \end{aligned}$$

Since $\dim\ker T > 0$, $\ker T$ contains some $v \in V \setminus \{\mathbf{0}\}$, so $\ker T \neq \{\mathbf{0}\}$. Hence T is not injective. □

No linear map from a finite-dimensional vector space to a "bigger" vector space can be surjective:

Proposition 4.13. *Suppose V and W are finite-dimensional vector spaces, $\dim V < \dim W$. Then there does not exist $T \in \mathscr{L}(V,W)$ such that T is surjective.*

Proof. Let $T \in \mathscr{L}(V,W)$. Then

$$\begin{aligned} \dim\operatorname{im} T &= \dim V - \dim\ker T && \text{[by fundamental theorem of linear maps]} \\ &\leq \dim V && [\because \dim\ker T \geq 0] \\ &< \dim W. \end{aligned}$$

Since $\dim\operatorname{im} T < \dim W$, $\operatorname{im} T \neq W$ so T is not surjective. □

Example (Homogeneous system of linear equations)**.** Consider the homogeneous system of linear equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= 0 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= 0 \end{aligned} \tag{$*$}$$

where $a_{ij} \in \mathbf{F}$.
Define $T : \mathbf{F}^n \to \mathbf{F}^m$ by

$$T(x_1,\ldots,x_n) = \left(\sum_{i=1}^{n} a_{1i}x_i, \ldots, \sum_{i=1}^{n} a_{mi}x_i\right).$$

The solution set of $(*)$ is given by

$$\ker T = \left\{(x_1,\ldots,x_n) \in \mathbf{F}^n \;\middle|\; \sum_{i=1}^{n} a_{1i}x_i = 0, \ldots, \sum_{i=1}^{n} a_{mi}x_i = 0\right\}.$$

Proposition. *A homogeneous system of linear equations with more variables than*

equations has non-zero solutions.

Proof. If $n > m$, then

$$\begin{aligned} \dim \mathbf{F}^n > \dim \mathbf{F}^m &\implies T \text{ is not injective} \\ &\implies \ker T \neq \{\mathbf{0}\} \\ &\implies (*) \text{ has non-zero solutions} \end{aligned}$$

□

Proposition. *A system of linear equations with more equations than variables has no solution for some choice of the constant terms.*

Proof. If $n < m$, then $\dim \mathbf{F}^n < \dim \mathbf{F}^m$, so T is not surjective. Hence there exists $(c_1, \ldots, c_m) \in \mathbf{F}^m$ such that

$$\forall (x_1, \ldots, x_n) \in \mathbf{F}^n, \quad T(x_1, \ldots, x_n) \neq (c_1, \ldots, c_m).$$

Thus the choice of constant terms $(c_1, \ldots, c_m)$ is such that the system of linear equations

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= c_1 \\ &\vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= c_m \end{aligned}$$

has no solutions $(x_1, \ldots, x_n)$. □

4.3 Matrices

Representing a Linear Map by a Matrix

Definition 4.14 (Matrix)**.** Suppose $m,n \in \mathbb{N}$. An $m \times n$ ***matrix*** A is a rectangular array with m rows and n columns:

$$A = \begin{pmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & & \vdots \\ A_{m1} & \cdots & A_{mn} \end{pmatrix}$$

where $A_{ij} \in \mathbf{F}$ denotes the entry in row i, column j.

Notation. We use i for indexing across the m rows, and j for indexing across the n columns.

Let $\mathscr{M}_{m\times n}(\mathbf{F})$ denotes the set of $m \times n$ matrices with entries in $\mathbf{F}$.

As we will soon see, matrices provide an efficient method of recording the values of Tv_j's in terms of a basis of W.

Definition 4.15 (Matrix of linear map)**.** Suppose $T \in \mathscr{L}(V,W)$, $\{v_1,\ldots,v_n\}$ is a basis of V, $\{w_1,\ldots,w_m\}$ is a basis of W. The ***matrix of*** T with respect to these bases is the $m \times n$ matrix $\mathscr{M}(T)$, whose entries A_{ij} are defined by

$$Tv_j = \sum_{i=1}^{m} A_{ij} w_i.$$

That is, the j-th column of $\mathscr{M}(T)$ consists of the scalars $A_{1j},\ldots,A_{mj}$ needed to write Tv_j as a linear combination of the bases of W.

Notation. If the bases of V and W are not clear from the context, we adopt the notation

$$\mathscr{M}(T;\{v_1,\ldots,v_n\},\{w_1,\ldots,w_m\}).$$

Addition and Scalar Multiplication of Matrices

Define addition and scalar multiplication on $\mathscr{M}_{m\times n}(\mathbf{F})$ as

$$(A+B)_{ij} = A_{ij} + B_{ij}$$
$$(\lambda A)_{ij} = \lambda A_{ij}$$

Lemma 4.16. *Suppose $S,T \in \mathscr{L}(V,W)$. Then*

(i) $\mathscr{M}(S+T) = \mathscr{M}(S) + \mathscr{M}(T)$;

(ii) $\mathscr{M}(\lambda T) = \lambda \mathscr{M}(T)$ *for* $\lambda \in \mathbf{F}$.

Proof. Suppose $S,T \in \mathscr{L}(V,W)$, $\{v_1,\ldots,v_n\}$ is a basis of V, $\{w_1,\ldots,w_m\}$ is a basis of W.

(i) Let $\mathscr{M}(S) = A$, $\mathscr{M}(T) = B$. Then

$$Sv_j = \sum_{i=1}^{m} A_{ij} w_i, \quad Tv_j = \sum_{i=1}^{m} B_{ij} w_i.$$

Let $\mathscr{M}(S+T) = C$. Then

$$\begin{aligned}
(S+T)v_j &= \sum_{i=1}^{m} C_{ij} w_i \\
Sv_j + Tv_j &= \sum_{i=1}^{m} C_{ij} w_i \\
\sum_{i=1}^{m} A_{ij} w_i + \sum_{i=1}^{m} B_{ij} w_i &= \sum_{i=1}^{m} C_{ij} w_i \\
\sum_{i=1}^{m} (A_{ij} + B_{ij}) w_i &= \sum_{i=1}^{m} C_{ij} w_i \\
A_{ij} + B_{ij} &= C_{ij}
\end{aligned}$$

which implies $A + B = C$. Hence $\mathscr{M}(S+T) = \mathscr{M}(S) + \mathscr{M}(T)$.

(ii) Let $\mathscr{M}(T) = A$. Then

$$Tv_j = \sum_{i=1}^{m} A_{ij} w_i.$$

Let $\lambda \in \mathbf{F}$, $\mathscr{M}(\lambda T) = B$. Then

$$\begin{aligned}
\lambda T v_j &= \sum_{i=1}^{m} B_{ij} w_i \\
\lambda \sum_{i=1}^{m} A_{ij} w_i &= \sum_{i=1}^{m} B_{ij} w_i \\
\lambda A_{ij} &= B_{ij}
\end{aligned}$$

which implies $\lambda A = B$. Hence $\mathscr{M}(\lambda T) = \lambda \mathscr{M}(T)$.

□

Lemma 4.17. *With addition and scalar multiplication defined as above, $\mathscr{M}_{m\times n}(\mathbf{F})$ is a vector space of dimension mn.*

Proof. The verification that $\mathscr{M}_{m\times n}(\mathbf{F})$ is a vector space is left to the reader. Note that the additive identity of $\mathscr{M}_{m\times n}(\mathbf{F})$ is the *zero matrix*; the $m\times n$ matrix all of whose entries equal 0.

The reader should also verify that the list of distinct $m\times n$ matrices that have 0 in all entries except for a 1 in one entry is a basis of $\mathscr{M}_{m\times n}(\mathbf{F})$. There are mn such matrices, so the dimension of $\mathscr{M}_{m\times n}(\mathbf{F})$ equals mn. □

Matrix Multiplication

Note that we define the product of two matrices only when the number of columns of the first matrix equals the number of rows of the second matrix.

Definition 4.18 (Matrix multiplication). Suppose $A\in\mathscr{M}_{m\times n}(\mathbf{F})$, $B\in\mathscr{M}_{n\times p}(\mathbf{F})$. Then $AB\in\mathscr{M}_{m\times p}$ is defined as

$$(AB)_{ij}=\sum_{k=1}^{n}A_{ik}B_{kj}.$$

This means that the entry in row i, column j of AB is computed by taking row i of A and column j of B, multiplying together corresponding entries, and then summing.

In the next result, we assume that the same basis of V is used in considering $T\in\mathscr{L}(U,V)$ and $S\in\mathscr{L}(V,W)$, the same basis of W is used in considering $S\in\mathscr{L}(V,W)$ and $ST\in\mathscr{L}(U,W)$, and the same basis of U is used in considering $T\in\mathscr{L}(U,V)$ and $ST\in\mathscr{L}(U,W)$.

Lemma 4.19 (Matrix of product of linear maps). *If $T\in\mathscr{L}(U,V)$ and $S\in\mathscr{L}(V,W)$, then $\mathscr{M}(ST)=\mathscr{M}(S)\mathscr{M}(T)$.*

Proof. Let $\{v_1,\dots,v_n\}$ be a basis of V, $\{w_1,\dots,w_m\}$ be a basis of W, $\{u_1,\dots,u_p\}$ be a basis of U.

Let $\mathscr{M}(S) = A$, $\mathscr{M}(T) = B$. For $j = 1, \ldots, p$,

$$\begin{aligned}(ST)u_j &= S(Tu_j) \\ &= S\left(\sum_{k=1}^{n} B_{kj} v_k\right) \\ &= \sum_{k=1}^{n} B_{kj} S v_k \\ &= \sum_{k=1}^{n} B_{kj} \left(\sum_{i=1}^{m} A_{ik} w_i\right) \\ &= \sum_{i=1}^{m} \left(\sum_{k=1}^{n} A_{ik} B_{kj}\right) w_i.\end{aligned}$$

□

Notation. Let $A_{i,\cdot}$ denote the row vector corresponding to the i-th row of A, and let $A_{\cdot,j}$ denote the column vector corresponding to the j-th column of A.

Lemma 4.20. *Suppose $A \in \mathscr{M}_{m\times n}(\mathbf{F})$, $B \in \mathscr{M}_{n\times p}(\mathbf{F})$. Then*

$$(AB)_{ij} = A_{i,\cdot} B_{\cdot,j}.$$

That is, the entry in row i, column j of AB equals (row i of A) times (column j of B).

Proof. By definition of matrix multiplication,

$$A_{i,\cdot} B_{\cdot,j} = \begin{pmatrix} A_{i1} & \cdots & A_{in} \end{pmatrix} \begin{pmatrix} B_{1j} \\ \vdots \\ B_{nj} \end{pmatrix} = \sum_{k=1}^{n} a_{ik} b_{kj} = (AB)_{ij}.$$

□

Lemma 4.21. *Suppose $A \in \mathscr{M}_{m\times n}(\mathbf{F})$, $B \in \mathscr{M}_{n\times p}(\mathbf{F})$. Then*

$$(AB)_{\cdot,j} = AB_{\cdot,j} \quad (j = 1, \ldots, p).$$

That is, column j of AB equals A times column j of B.

Proof. Using the previous result,

$$AB_{\cdot,j} = \begin{pmatrix} A_{1,\cdot} B_{\cdot,j} \\ \vdots \\ A_{n,\cdot} B_{\cdot,j} \end{pmatrix} = \begin{pmatrix} (AB)_{1j} \\ \vdots \\ (AB)_{nj} \end{pmatrix} = (AB)_{\cdot,j}$$

□

Lemma 4.22 (Linear combination of columns). *Suppose* $A \in \mathscr{M}_{m\times n}(\mathbf{F})$, $b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$.

Then

$$Ab = b_1 A_{\cdot,1} + \cdots + b_n A_{\cdot,n}.$$

That is, Ab is a linear combination of the columns of A, with the scalars that multiply the columns coming from b.

Proof. We have

$$Ab = \begin{pmatrix} A_{11}b_1 + \cdots + A_{1n}b_n \\ \vdots \\ A_{m1}b_1 + \cdots + A_{mn}b_n \end{pmatrix} = \begin{pmatrix} A_{11}b_1 \\ \vdots \\ A_{m1}b_1 \end{pmatrix} + \cdots + \begin{pmatrix} A_{1n}b_n \\ \vdots \\ A_{mn}b_n \end{pmatrix}$$
$$= b_1 \begin{pmatrix} A_{11} \\ \vdots \\ A_{m1} \end{pmatrix} + \cdots + b_n \begin{pmatrix} A_{1n} \\ \vdots \\ A_{mn} \end{pmatrix} = b_1 A_{\cdot,1} + \cdots + b_n A_{\cdot,n}.$$

□

The next result is the main tool used to prove the column–row factorisation 4.27 and to prove that the column rank of a matrix equals the row rank. To be consistent with the notation often used with the column–row factorisation, we denote the matrices as C and R instead of A and B.

Lemma 4.23. *Suppose* $C \in \mathscr{M}_{m\times c}(\mathbf{F})$, $R \in \mathscr{M}_{c\times n}(\mathbf{F})$. *Then*

(i) Columns: for $j = 1,\ldots,n$, $(CR)_{\cdot,j}$ *is a linear combination of* $C_{\cdot,1},\ldots,C_{\cdot,c}$, *with coefficients coming from* $R_{\cdot,j}$.

(ii) Rows: for $i = 1,\ldots,m$, $(CR)_{i,\cdot}$ *is a linear combination of* $R_{1,\cdot},\ldots,R_{c,\cdot}$, *with coefficients coming from* $C_{j,\cdot}$.

Proof.

(i) Suppose $j \in \{1,\ldots,n\}$.

$$(CR)_{\cdot,j} = CR_{\cdot,j}$$

and then apply the previous result.

(ii) Similar.

□

Rank of a Matrix

Definition 4.24. Suppose $A \in \mathscr{M}_{m\times n}(\mathbf{F})$. The *row space* of A is the span of its rows, and the *column space* of A is the span of its columns:

$$\begin{aligned}\mathrm{Row}(A) &:= \mathrm{span}\,(A_{i,\cdot} \mid 1 \le i \le m)\,,\\ \mathrm{Col}(A) &:= \mathrm{span}\,\big(A_{\cdot,j} \mid 1 \le j \le n\big)\,.\end{aligned}$$

The ***row rank*** and ***column rank*** of A are defined as

$$\begin{aligned}r(A) &:= \dim \mathrm{Row}(A),\\ c(A) &:= \dim \mathrm{Col}(A).\end{aligned}$$

If A is an $m \times n$ matrix, then the column rank of A is at most n (because A has n columns) and the column rank of A is also at most m (because $\dim \mathscr{M}_{m\times 1} = m$). Similar remarks hold for the row rank of A.

We now define the *transpose* of a matrix.

Definition 4.25 (Transpose)**.** Suppose $A \in \mathscr{M}_{m\times n}(\mathbf{F})$. The ***transpose*** of A is the $n \times m$ matrix $A^\top$ whose entries are defined by

$$(A^\top)_{ij} = A_{ji}.$$

Lemma 4.26 (Properties of transpose)**.** *Suppose $A, B \in \mathscr{M}_{m\times n}(\mathbf{F})$, $C \in \mathscr{M}_{n\times p}(\mathbf{F})$.*

(i) $(A+B)^\top = A^\top + B^\top$.

(ii) $(\lambda A)^\top = \lambda A^\top$ *for* $\lambda \in \mathbf{F}$.

(iii) $(AC)^\top = C^\top A^\top$.

Proof.

(i) $(A+B)^\top{}_{ij} = (A+B)_{ji} = A_{ji} + B_{ji} = (A^\top)_{ij} + (B^\top)_{ij}$

(ii) $(\lambda A)^\top{}_{ij} = (\lambda A)_{ji} = \lambda A_{ji} = \lambda (A^\top)_{ij}$

(iii) $(AC)^\top{}_{ij} = (AC)_{ji} = \sum_{k=1}^n A_{jk}C_{ji} = \sum_{k=1}^n C_{ji}A_{jk} = \sum_{k=1}^n (C^\top)_{ik}(A^\top)_{kj} = (C^\top A^\top)_{ij}$

□

The next result will be the main tool used to prove that the column rank equals the row rank.

Proposition 4.27 (Column-row factorisation). *Suppose $A \in \mathscr{M}_{m\times n}(\mathbf{F})$, $c(A) \geq 1$. Then there exist $C \in M_{m\times c(A)}(\mathbf{F})$, $R \in M_{c(A)\times n}(\mathbf{F})$ such that $A = CR$.*

Proof. We prove by construction, i.e., construct the required matrices C and R.

Each column of A is a $m \times 1$ matrix. The set of columns of A

$$\{A_{\cdot,1}, \ldots, A_{\cdot,n}\}$$

is a spanning set of $\mathrm{Col}(A)$, so it can be reduced to a basis of $\mathrm{Col}(A)$, by 3.25. This basis has length $c(A)$, by the definition of column rank.

The $c(A)$ columns in this basis can be put together to form a $m \times c(A)$ matrix, which we call C.

For $j \in \{1, \ldots, n\}$, the j-th column of A is a linear combination of the columns of C. Make the coefficients of this linear combination into column j of a $c(A) \times n$ matrix, which we call R. By 4.23(i), it follows that $A = CR$. □

Theorem 4.28. *The column rank of a matrix equals its row rank.*

Proof. Suppose $A \in \mathscr{M}_{m\times n}(\mathbf{F})$. Let $A = CR$ be the column-row factorisation of A given by 4.27, where $C \in \mathscr{M}_{m\times c(A)}(\mathbf{F})$, $R \in \mathscr{M}_{c(A)\times n}(\mathbf{F})$.

Then 4.23(ii) tells us that every row of A is a linear combination of the rows of R. Because R has $c(A)$ rows, this implies that the row rank of A is less than or equal to the column rank $c(A)$ of A.

To prove the inequality in the other direction, apply the result in the previous paragraph to $A^\top$, getting

$$\begin{aligned} c(A) &= r(A^\top) \\ &\leq c(A^\top) \\ &= r(A). \end{aligned}$$

Thus the column rank of A equals the row rank of A. □

Since the column rank equals row rank, we can dispense with the terms "column rank" and "row rank", and just use the simpler term "rank".

Definition 4.29 (Rank). The ***rank*** of a matrix A is defined as

$$\mathrm{rank}\, A := r(A) = c(A).$$

4.4 Invertibility and Isomorphism

Invertibility

Definition 4.30 (Invertibility). We say $T \in \mathscr{L}(V,W)$ is ***invertible*** if there exists $S \in \mathscr{L}(W,V)$ such that $ST = I_V$, $TS = I_W$; we call S an *inverse* of T.

The next result allows us to talk about "the" inverse of a linear map.

Lemma. *The inverse of an invertible linear map is unique.*

Proof. Suppose $T \in \mathscr{L}(V,W)$ is invertible, $S_1, S_2 \in \mathscr{L}(W,V)$ are inverses of T. Then

$$S_1 = S_1 I_W = S_1(TS_2) = (S_1T)S_2 = I_V S_2 = S_2.$$

□

Since the inverse is unique, we can give it a notation.

Notation. If T is invertible, we denote its inverse by T^{-1}.

The following result is useful in determing if a linear map is invertible.

Lemma 4.31 (Invertibility criterion). *Suppose $T \in \mathscr{L}(V,W)$.*

(i) T is invertible $\iff$ T is injective and surjective.

(ii) If $\dim V = \dim W$, T is invertible $\iff$ T is injective $\iff$ T is surjective.

Proof.

(i) $\boxed{\Longrightarrow}$ Suppose $T \in \mathscr{L}(V,W)$ is invertible with inverse T^{-1}.

Suppose $Tu = Tv$. Applying T^{-1} to both sides gives $u = v$. Hence T is injective.

Let $w \in W$. Then $w = T\left(T^{-1}w\right)$, which shows that $w \in \operatorname{im} T$, so $\operatorname{im} T = W$. Hence T is surjective.

$\boxed{\Longleftarrow}$ Suppose T is injective and surjective.

Define $S \in \mathscr{L}(W,V)$ such that for each $w \in W$, $S(w)$ is the unique element of V such that $T(S(w)) = w$ (we can do this due to injectivity and surjectivity). Then we have that $T(ST)v = (TS)Tv = Tv$ and thus $STv = v$ so $ST = I$. It is easy to show that S is a linear map.

(ii) It suffices to only prove T is injective $\iff$ T is surjective. Then apply the previous result.

$\boxed{\implies}$ Suppose T is injective. Then $\ker T = \{\mathbf{0}\}$, so $\dim\ker T = 0$. By the fundamental theorem of linear maps,

$$\dim\operatorname{im} T = \dim V - \dim\ker T = \dim V = \dim W$$

which implies that T is surjective.

$\boxed{\impliedby}$ Suppose T is surjective. Then $\dim\operatorname{im} T = \dim W$. By the fundamental theorem of linear maps,

$$\dim\ker T = \dim V - \dim\operatorname{im} T = \dim V - \dim W = 0$$

which implies that T is injective.

□

Corollary 4.32. *Suppose V and W are finite-dimensional, $\dim V = \dim W$, $S \in \mathscr{L}(W,V)$, $T = \mathscr{L}(V,W)$. Then $ST = I$ if and only if $TS = I$.*

Proof.

$\boxed{\implies}$ Suppose $ST = I$. Let $v \in \ker T$. Then

$$v = Iv = (ST)v = S(Tv) = S(\mathbf{0}) = \mathbf{0} \implies \ker T = \{\mathbf{0}\}$$

so T is injective. Since $\dim V = \dim W$, by 4.31, T is invertible.

Since $ST = I$, then

$$S = STT^{-1} = IT^{-1} = T^{-1}$$

so $TS = TT^{-1} = I$, as desired.

$\boxed{\impliedby}$ Similar to the above; reverse the roles of S and T (and V and W) to show that if $TS = I$ then $ST = I$. □

Isomorphism

The next definition captures the idea of two vector spaces that are essentially the same, except for the names of their elements.

Definition 4.33. An ***isomorphism*** is an invertible linear map.
We say V is ***isomorphic*** to W, denote $V \cong W$, if there exists an isomorphism $T \in \mathscr{L}(V,W)$.

The next result shows that we need to look at only at the dimension to determine whether two vector spaces are isomorphic.

Lemma 4.34. *Suppose V and W are finite-dimensional. Then*

$$V \cong W \iff \dim V = \dim W.$$

Proof.

$\boxed{\Longrightarrow}$ Suppose $V \cong W$. Then there exists an isomorphism $T \in \mathscr{L}(V,W)$, which is invertible. By 4.31, T is both injective and surjective. Thus $\ker T = \{\mathbf{0}\}$ and $\operatorname{im} T = W$, implying $\dim \ker T = 0$ and $\dim \operatorname{im} T = \dim W$.

By the fundamental theorem of linear maps,

$$\begin{aligned} \dim V &= \dim \ker T + \dim \operatorname{im} T \\ &= 0 + \dim W = \dim W. \end{aligned}$$

$\boxed{\Longleftarrow}$ Suppose V and W are finite-dimensional, $\dim V = \dim W = n$. Let $\{v_1, \ldots, v_n\}$ be a basis of V, $\{w_1, \ldots, w_n\}$ be a basis of W.

It suffices to construct an surjective $T \in \mathscr{L}(V,W)$. By the linear map lemma, there exists a linear map $T \in \mathscr{L}(V,W)$ such that

$$Tv_i = w_i \quad (i = 1, \ldots, n)$$

Let $w \in W$. Then there exist $a_i \in \mathbf{F}$ such that $w = a_1 w_1 + \cdots + a_n w_n$. Then

$$\begin{aligned} T(a_1 v_1 + \cdots + a_n v_n) = w &\implies w \in \operatorname{im} T \\ &\implies W = \operatorname{im} T \\ &\implies T \text{ is surjective} \\ &\implies T \text{ is invertible.} \end{aligned}$$

□

Proposition 4.35. *Suppose $\{v_1, \ldots, v_n\}$ is a basis of V, $\{w_1, \ldots, w_m\}$ is a basis of W. Then*

$$\mathscr{L}(V,W) \cong \mathscr{M}_{m \times n}(\mathbf{F}).$$

Proof. We claim that $\mathscr{M}$ is an isomorphism between $\mathscr{L}(V,W)$ and $\mathscr{M}_{m \times n}(\mathbf{F})$.

We already noted that $\mathscr{M}$ is linear. We need to prove that $\mathscr{M}$ is (i) injective and (ii) surjective.

(i) Given $T \in \mathscr{L}(V,W)$, if $\mathscr{M}(T) = 0$, then

$$Tv_j = 0 \quad (j = 1,\dots,n)$$

Since $v_1,\dots,v_n$ is a basis of V, this implies $T = \mathbf{0}$, so $\ker \mathscr{M} = \{\mathbf{0}\}$. Thus $\mathscr{M}$ is injective.

(ii) Suppose $A \in \mathscr{M}_{m\times n}(\mathbf{F})$. By the linear map lemma, there exists $T \in \mathscr{L}(V,W)$ such that

$$Tv_j = \sum_{i=1}^{m} A_{ij} w_i \quad (j = 1,\dots,n)$$

Since $\mathscr{M}(T) = A$, $\operatorname{im} \mathscr{M} = \mathscr{M}_{m\times n}(\mathbf{F})$ so $\mathscr{M}$ is surjective.

□

Now we can determine the dimension of the vector space of linear maps from one finite-dimensional vector space to another.

Corollary 4.36. *Suppose V and W are finite-dimensional. Then $\mathscr{L}(V,W)$ is finite-dimensional and*

$$\dim \mathscr{L}(V,W) = (\dim V)(\dim W).$$

Proof. Since $\mathscr{L}(V,W) \cong \mathscr{M}_{m\times n}(\mathbf{F})$,

$$\dim \mathscr{L}(V,W) = \dim \mathscr{M}_{m\times n}(\mathbf{F}) = mn = (\dim V)(\dim W).$$

□

Linear Maps Thought of as Matrix Multiplication

Previously we defned the matrix of a linear map. Now we defne the matrix of a vector.

Definition 4.37 (Matrix of a vector)**.** Suppose $v \in V$, $\{v_1,\dots,v_n\}$ is a basis of V. The matrix of v with respect to this basis is

$$\mathscr{M}(v) = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

where $b_1,\dots,b_n \in \mathbf{F}$ are such that

$$v = b_1 v_1 + \cdots + b_n v_n.$$

Example. If $x = (x_1, \ldots, x_n) \in \mathbf{F}^n$, then the matrix of the vector x with respect to the standard basis of $\mathbf{F}^n$ is

$$\mathscr{M}(x) = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

Lemma 4.38. *Suppose $T \in \mathscr{L}(V,W)$. Let $\{v_1, \ldots, v_n\}$ be a basis of V, $\{w_1, \ldots, w_m\}$ be a basis of W. Then*

$$\mathscr{M}(T)_{\cdot,j} = \mathscr{M}(Tv_j) \quad (j = 1, \ldots, n)$$

Proof. By definition, the entries of $\mathscr{M}(T)$ are defined such that

$$Tv_j = \sum_{i=1}^{m} A_{ij} w_i \quad (j = 1, \ldots, n)$$

Then since $Tv_j \in W$, by definition, the matrix of Tv_j with respect to the basis $\{w_1, \ldots, w_m\}$ is

$$\mathscr{M}(Tv_j) = \begin{pmatrix} A_1 j \\ \vdots \\ A_{mj} \end{pmatrix}$$

which is precisely the j-th column of $\mathscr{M}(T)_{\cdot,j}$. □

The following result shows that linear maps act like matrix multiplication.

Lemma 4.39. *Suppose $T \in \mathscr{L}(V,W)$. Let $\{v_1, \ldots, v_n\}$ be a basis of V, $\{w_1, \ldots, w_m\}$ be a basis of W. Let $v \in V$, then*

$$\mathscr{M}(Tv) = \mathscr{M}(T)\mathscr{M}(v).$$

Proof. Suppose $v = b_1 v_1 + \cdots + b_n v_n$ for some $b_1, \ldots, b_n \in \mathbf{F}$. Then

$$\begin{aligned} \mathscr{M}(Tv) &= \mathscr{M}\left(T(b_1 v_1 + \cdots + b_n v_n)\right) \\ &= b_1 \mathscr{M}(Tv_1) + \cdots + b_n \mathscr{M}(Tv_n) \\ &= b_1 \mathscr{M}(T)_{\cdot,1} + \cdots + b_n \mathscr{M}(T)_{\cdot,n} \\ &= \begin{pmatrix} \mathscr{M}(T)_{\cdot,1} & \cdots & \mathscr{M}(T)_{\cdot,n} \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \\ &= \mathscr{M}(T)\mathscr{M}(v). \end{aligned}$$

□

Notice that no bases are in sight in the statement of the next result. Although $\mathscr{M}(T)$ in the next result depends on a choice of bases of V and W, the next result shows that the column rank of $\mathscr{M}(T)$ is the same for all such choices (because $\operatorname{im} T$ does not depend on a choice of basis).

Proposition 4.40. *Suppose V and W are finite-dimensional, $T \in \mathscr{L}(V,W)$. Then*

$$\dim \ker T = \operatorname{rank} \mathscr{M}(T).$$

Proof. Suppose $\{v_1,\dots,v_n\}$ is a basis of V, $\{w_1,\dots,w_m\}$ is a basis of W.

The linear map that takes $w \in W$ to $\mathscr{M}(w)$ is an isomorphism from W to $\mathscr{M}_{m\times 1}(\mathbf{F})$ (consisting of $m \times 1$ column vectors).

The restriction of this isomorphism to $\operatorname{im} T$ [which equals $\operatorname{span}(Tv_1,\dots,Tv_n)$] is an isomorphism from $\operatorname{im} T$ to $\operatorname{span}(\mathscr{M}(Tv_1),\dots,\mathscr{M}(Tv_n))$. For $j = 1,\dots,n$, the $m \times 1$ matrix $\mathscr{M}(Tv_j)$ equals column k of $\mathscr{M}(T)$. Thus

$$\dim \ker T = \operatorname{rank} \mathscr{M}(T),$$

as desired. □

Change of Basis

For $n \in \mathbb{N}$, the $n \times n$ ***identity matrix*** is

$$I_n = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}.$$

Remark. The symbol I is used to denote both the identity operator and the identity matrix. The context indicates which meaning of I is intended.

The next result justifies the name "identity matrix".

Lemma 4.41. *Suppose $A \in \mathscr{M}_{n\times n}(\mathbf{F})$. Then $AI_n = I_n A = A$.*

Proof. Exercise. □

Definition 4.42 (Invertible matrix). We say $A \in \mathscr{M}_{n\times n}(\mathbf{F})$ is ***invertible*** if there exists $B \in \mathscr{M}_{n\times n}(\mathbf{F})$ such that $AB = BA = I$; we call B an *inverse* of A.

Lemma. *The inverse of an invertible square matrix is unique.*

Proof. Let A be an invertible square matrix, let B and C be inverses of A. Then

$$B = BI = BAC = IC = C.$$

□

Since the inverse of a matrix is unique, we can give it a notation.

Notation. The inverse of a matrix A is denoted by A^{-1}.

Lemma 4.43.

(i) Suppose A is an invertible square matrix. Then $(A^{-1})^{-1} = A$.

(ii) Suppose A and C are invertible square matrices of the same size. Then AC is invertible, and $(AC)^{-1} = C^{-1}A^{-1}$.

Proof.

(i) We have

$$A^{-1}A = AA^{-1} = I,$$

so the inverse of A^{-1} is A.

(ii) We have

$$\begin{aligned}(AC)(C^{-1}A^{-1}) &= A(CC^{-1})A^{-1}\\ &= AIA^{-1}\\ &= AA^{-1}\\ &= I,\end{aligned}$$

and similarly $(C^{-1}A^{-1})(AC) = I$.

□

Lemma 4.44 (Matrix of product of linear maps). *Suppose $T \in \mathscr{L}(U,V)$, $S \in \mathscr{L}(V,W)$. Let $\{u_1,\dots,u_m\}$ be a basis of U, $\{v_1,\dots,v_n\}$ be a basis of V, $\{w_1,\dots,w_p\}$ be a basis of W. Then*

$$\begin{aligned}&\mathscr{M}(ST;\{u_1,\dots,u_m\},\{w_1,\dots,w_p\}) =\\ &\qquad \mathscr{M}(S;\{v_1,\dots,v_n\},\{w_1,\dots,w_p\})\,\mathscr{M}(T;\{u_1,\dots,u_m\},\{v_1,\dots,v_n\}).\end{aligned}$$

Proof. Refer to previous section. Now we are just being more explicit about the bases involved.

□

Corollary 4.45. *Let $\{u_1,\dots,u_n\}$ and $\{v_1,\dots,v_n\}$ be bases of V. Then the matrices*

$$\mathscr{M}(I;\{u_1,\dots,u_n\},\{v_1,\dots,v_n\}) \quad \text{and} \quad \mathscr{M}(I;\{v_1,\dots,v_n\},\{u_1,\dots,u_n\})$$

are invertible, and each is the inverse of the other.

Proof. In the previous result, replace w_i with u_i, and replace S and T with I, to obtain

$$I = \mathscr{M}(I;\{v_1,\dots,v_n\},\{u_1,\dots,u_n\})\mathscr{M}(I;\{u_1,\dots,u_n\},\{v_1,\dots,v_n\}).$$

Now interchange the roles of u's and v's, which gives

$$I = \mathscr{M}(I;\{u_1,\dots,u_n\},\{v_1,\dots,v_n\})\mathscr{M}(I;\{v_1,\dots,v_n\},\{u_1,\dots,u_n\}).$$

These two equations above give the desired result. □

Proposition 4.46 (Change-of-basis formula)**.** *Suppose $T \in \mathscr{L}(V)$. Let $\{u_1,\dots,u_n\}$ and $\{v_1,\dots,v_n\}$ be bases of V. Let*

$$A = \mathscr{M}(T;\{u_1,\dots,u_n\}), \quad B = \mathscr{M}(T;\{v_1,\dots,v_n\}),$$

and $C = \mathscr{M}(I;\{u_1,\dots,u_n\},\{v_1,\dots,v_n\})$. Then

$$A = C^{-1}BC. \tag{4.2}$$

Proof. Note that

$$\begin{aligned}\mathscr{M}(T;\{u_1,\dots,u_n\},\{v_1,\dots,v_n\}) &= \underbrace{\mathscr{M}(T;\{v_1,\dots,v_n\})}_{B}\underbrace{\mathscr{M}(I;\{u_1,\dots,u_n\},\{v_1,\dots,v_n\})}_{C} \\ &= \underbrace{\mathscr{M}(I;\{u_1,\dots,u_n\},\{v_1,\dots,v_n\})}_{C}\underbrace{\mathscr{M}(T;\{u_1,\dots,u_n\})}_{A}\end{aligned}$$

Hence $BC = CA$, and the desired result follows. □

THe next result states that the matrix of inverse equals the inverse of matrix.

Lemma 4.47. *Suppose $\{v_1,\dots,v_n\}$ is a basis of V, $T \in \mathscr{L}(V)$ is invertible. Then*

$$\mathscr{M}\left(T^{-1}\right) = (\mathscr{M}(T))^{-1},$$

where both matrices are with respect to the basis $\{v_1,\dots,v_n\}$.

Proof. We have that

$$\mathscr{M}\left(T^{-1}\right)\mathscr{M}(T) = \mathscr{M}\left(T^{-1}T\right) = \mathscr{M}(I) = I.$$

□

4.5 Products and Quotients of Vector Spaces

Products of Vector Spaces

As usual when dealing with more than one vector space, all vector spaces in use should be over the same field.

Definition 4.48 (Product)**.** Suppose $V_1, \ldots, V_n$ are vector spaces over $\mathbf{F}$. The ***product*** $V_1 \times \cdots \times V_n$ is defined by

$$V_1 \times \cdots \times V_n := \{(v_1, \ldots, v_n) \mid v_i \in V_i\}.$$

Remark. This is analagous to the Cartesian product of sets.

Lemma. *$V_1 \times \cdots \times V_n$ is a vector space over $\mathbf{F}$, with addition and scalar multiplication defined by*

$$(u_1, \ldots, u_n) + (v_1, \ldots, v_n) = (u_1 + v_1, \ldots, u_n + v_n)$$
$$\lambda(v_1, \ldots, v_n) = (\lambda v_1, \ldots, \lambda v_n)$$

Proof. The proof is left as an exercise; the additive identity is $\mathbf{0} = (0, \ldots, 0)$, and the additive inverse of $(v_1, \ldots, v_n)$ is $(-v_1, \ldots, -v_n)$. □

The next result shows that the dimension of a product is the sum of dimensions.

Lemma 4.49 (Dimension of product)**.** *Suppose $V_1, \ldots, V_n$ are finite-dimensional. Then $V_1 \times \cdots \times V_n$ is finite-dimensional, and*

$$\dim(V_1 \times \cdots \times V_n) = \dim V_1 + \cdots + \dim V_n.$$

Proof. Choose a basis of each V_i. For each basis vector of each V_i, consider the element of $V_1 \times \cdots \times V_n$ that equals the basis vector in the i-th slot and 0 in the other slots. The set of all such vectors is linearly independent and spans $V_1 \times \cdots \times V_n$. Thus it is a basis of $V_1 \times \cdots \times V_n$. The length of this basis is $\dim V_1 + \cdots + \dim V_n$, as desired.

□

The next result says that a sum is a direct sum if and only if dimensions add up.

Proposition 4.50. *Suppose V is finite-dimensional, $V_1, \ldots, V_n \leq V$. Then $V_1 + \cdots + V_n$ is a direct sum if and only if*

$$\dim(V_1 + \cdots + V_n) = \dim V_1 + \cdots + \dim V_n.$$

Proof. Define the map

$$\begin{aligned} \Gamma\colon V_1 \times \cdots \times V_n &\to V_1 + \cdots + V_n \\ (v_1, \ldots, v_n) &\mapsto v_1 + \cdots + v_n \end{aligned}$$

Γ is a linear map, as you should verify. Then

$$\begin{aligned} \Gamma \text{ is injective} &\iff \ker \Gamma = \{\mathbf{0}\} \\ &\iff (v_1, \ldots, v_n) = \mathbf{0} \\ &\iff v_1 = \cdots = v_n = 0 \\ &\iff V_1 \oplus \cdots \oplus V_n \qquad \text{[by 3.12]} \end{aligned}$$

The map Γ is surjective. Hence by the fundamental theorem of linear maps, Γ is injective if and only if

$$\begin{aligned} \dim(V_1 \times \cdots \times V_n) &= \dim \ker \Gamma + \dim \operatorname{im} \Gamma \\ &= 0 + \dim(V_1 + \cdots + V_n) \\ &= \dim(V_1 + \cdots + V_n). \end{aligned}$$

By 4.49, we are done. □

Powers of Vector Spaces

For a positive integer n, define the power of the vector space V as

$$V^n = \underbrace{V \times \cdots \times V}_{n \text{ times}}.$$

Proposition 4.51. $V^n \cong \mathscr{L}(\mathbf{F}^n, V)$.

Quotient Spaces

We begin our approach to quotient spaces by defining a *coset*.

Definition 4.52 (Coset). Suppose $v \in V$, $U \subset V$. We call $v + U$ a ***coset*** of U, defined by

$$v + U := \{v + u \mid u \in U\}.$$

Definition 4.53 (Quotient space)**.** Suppose $U \leq V$. Then the ***quotient space*** V/U is the set of cosets of U:

$$V/U := \{v+U \mid v \in V\}.$$

Example. Let $U = \{(x,2x) \in \mathbb{R}^2 \mid x \in \mathbb{R}\}$. Then $\mathbb{R}^2/U$ is the set of lines in $\mathbb{R}^2$ that have gradient of 2.

The next result shows that two cosets of a subspace are equal or disjoint.

Lemma 4.54. *Suppose $U \leq V$, and $v,w \in V$. Then*

$$v-w \in U \iff v+U = w+U \iff (v+U)\cap(w+U) \neq \emptyset.$$

Proof.

(i) $\implies$ (ii) Suppose $v-w \in U$.

If $u \in U$, then

$$v+u = w+((v-w)+u) \in w+U.$$

Thus $v+U \subset w+U$. Similarly, $w+U \subset v+U$. Thus $v+U = w+U$.

(ii) $\implies$ (iii) The equation $v+U = w+U$ implies that $(v+U)\cap(w+U) \neq \emptyset$.

(iii) $\implies$ (i) Suppose $(v+U)\cap(w+U) \neq \emptyset$. Thus there exist $u_1,u_2 \in U$ such that

$$v+u_1 = w+u_2.$$

Thus $v-w = u_2-u_1$. Hence $v-w \in U$. □

We can define a vector space structure on V/U.

Lemma 4.55. *Suppose $U \leq V$. Then V/U is a vector space, with addition and scalar multiplication defined by*

$$(v+U)+(w+U) = (v+w)+U$$
$$\lambda(v+U) = (\lambda v)+U$$

for all $v,w \in V$, $\lambda \in \mathbf{F}$.

Proof. We first show that addition and scalar multiplication are well-defined.

Addition. Let $v_1,v_2,w_1,w_2 \in V$ be such that

$$v_1+U = v_2+U, \quad w_1+U = w_2+U.$$

By 4.54, $v_1 - v_2 \in U$ and $w_1 - w_2 \in U$. Since U is closed under addition, so $(v_1 - v_2) + (w_1 - w_2) \in U$. Thus $(v_1 + w_1) - (v_2 + w_2) \in U$. Using 4.54 again, we have

$$(v_1 + w_1) + U = (v_2 + w_2) + U,$$

as desired. Hence addition on V/U is well-defined.

Scalar multiplication. Let $v_1, v_2 \in V$ be such that $v_1 + U = v_2 + U$, and let $\lambda \in \mathbf{F}$.

Since U is closed under scalar multiplication, $\lambda(v_1 - v_2) \in U$. Thus $\lambda v_1 - \lambda v_2 \in U$. By 4.54,

$$(\lambda v_1) + U = (\lambda v_2) + U.$$

Hence scalar multiplication on V/U is well-defined.

The verification that addition and scalar multiplication make V/U into a vector space is straightforward and is left to the reader; the additive identity of V/U is $0 + U = U$, and the additive inverse of $v + U$ is $(-v) + U$. □

We now define a natural (or *canonical*) map from a vector space to its quotient space.

Definition 4.56 (Quotient map). Suppose $U \leq V$. The ***quotient map*** is the map

$$\begin{aligned} \pi \colon V &\to V/U \\ v &\mapsto v + U \end{aligned}$$

for all $v \in V$.

Notation. Although π depends on U as well as V, these spaces are left out of the notation because they should be clear from the context.

Lemma. *The quotient map is a linear map.*

Proof. Let $v, w \in V$, $\lambda \in \mathbf{F}$.

(i) $\pi(v) + \pi(w) = (v + U) + (w + U) = (v + w) + U = \pi(v + w)$.

(ii) $\pi(\lambda v) = (\lambda v) + U = \lambda(v + U) = \lambda(\pi v)$.

□

The kernel of the quotient map is U:

$$\begin{aligned} v \in \ker \pi &\iff \pi(v) = \mathbf{0} + U = U \\ &\iff v + U = U \quad [\text{by } 4.54] \\ &\iff v \in U \end{aligned}$$

Hence $\ker \pi = U$.

The definition of the quotient map implies it is surjective. Hence $\operatorname{im} \pi = V/U$.

Lemma 4.57 (Dimension of quotient space). *Suppose V is finite-dimensional, $U \leq V$. Then*

$$\dim V/U = \dim V - \dim U.$$

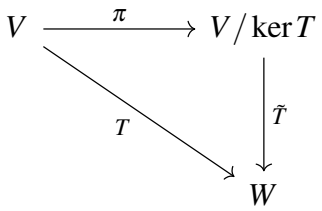
Proof. This follows from the fundamental theorem of linear maps:

$$\begin{aligned} \dim V &= \dim \ker \pi + \dim \operatorname{im} \pi \\ &= \dim U + \dim V/U. \end{aligned}$$

□

Suppose $T \in \mathscr{L}(V,W)$. Define the *induced map*

$$\begin{aligned} \tilde{T} \colon V/\ker T &\to W \\ v + \ker T &\mapsto Tv \end{aligned}$$



We first show that $\tilde{T}$ is well-defined.

Proof. Suppose $u, v \in V$ are such that $u + \ker T = v + \ker T$.

By 4.54, $u - v \in \ker T$. Thus $T(u-v) = \mathbf{0}$, so $Tu = Tv$. □

Lemma. *$\tilde{T}$ is a linear map from $V/\ker T$ to W.*

Proof. Let $u, v \in V$, $\lambda \in \mathbf{F}$.

(i) $\tilde{T}((u+\ker T)+(v+\ker T)) = \tilde{T}(u+v+\ker T) = T(u+v) = Tu+Tv = \tilde{T}(u+\ker T)+\tilde{T}(v+\ker T)$.

(ii) $\tilde{T}(\lambda(v+\ker T)) = \tilde{T}(\lambda v+\ker T) = T(\lambda v) = \lambda Tv = \lambda \tilde{T}(v+\ker T)$.

□

Lemma. *Suppose $T \in \mathscr{L}(V,W)$. Then $\tilde{T} \circ \pi = T$.*

Proof. Let $v \in V$. Then $(\tilde{T} \circ \pi)(v) = \tilde{T}(\pi(v)) = \tilde{T}(v+\ker T) = Tv$. □

Theorem 4.58 (First isomorphism theorem)**.** *Suppose $T \in \mathscr{L}(V,W)$. Then*

$$V/\ker T \cong \operatorname{im} T. \tag{4.3}$$

Proof.

Claim. $\tilde{T}$ is an isomorphism from $V/\ker T$ onto $\operatorname{im} T$.

- Let $v + \ker T \in \ker \tilde{T}$. Then

$$\begin{aligned} \tilde{T}(v+\ker T) = \mathbf{0} &\implies Tv = \mathbf{0} \\ &\implies v \in \ker T \\ &\implies v + \ker T = \ker T \end{aligned}$$

  so $\ker \tilde{T} = \{\mathbf{0} + \ker T\}$. Hence $\tilde{T}$ is injective.

- The definition of $\tilde{T}$ shows that $\operatorname{im} \tilde{T} = \operatorname{im} T$.

□

Theorem 4.59 (Second isomorphism theorem)**.** *Suppose $U, W \leq V$. Then*

$$(U+W)/W \cong U/(U \cap W). \tag{4.4}$$

Theorem 4.60 (Third isomorphism theorem)**.** *Suppose $U \subset V \subset W$. Then*

$$W/V \cong (W/U)/(V/U). \tag{4.5}$$

4.6 Duality

Dual Space and Dual Map

Linear maps into the scalar field $\mathbf{F}$ get a special name.

Definition 4.61 (Linear functional)**.** A ***linear functional*** on V is a linear map from V to $\mathbf{F}$.

That is, a linear functional is an element of $\mathscr{L}(V,\mathbf{F})$.

Example. $\phi : \mathbb{R}^3 \to \mathbb{R}$ defined by $\phi(x,y,z) = x+y+z$ is a linear functional on $\mathbb{R}^3$.

Definition 4.62 (Dual space)**.** The ***dual space*** V' of V is the vector space of linear functionals on V.

That is, $V' := \mathscr{L}(V,\mathbf{F})$.

Lemma 4.63 (Dimension of dual space)**.** *Suppose V is finite-dimensional. Then V' is finite-dimensional, and*

$$\dim V' = \dim V.$$

Proof. By 4.36,

$$\dim V' := \dim \mathscr{L}(V,\mathbf{F}) = (\dim V)(\dim \mathbf{F}) = \dim V.$$

□

Definition 4.64 (Dual basis)**.** Let $\{v_1,\dots,v_n\}$ be a basis of V. Then the ***dual basis*** of $\{v_1,\dots,v_n\}$ is

$$\{\phi_1,\dots,\phi_n\} \subset V',$$

where each ϕ_i is the linear functional on V such that

$$\phi_i(v_j) = \delta_{ij} = \begin{cases} 1 & (i=j) \\ 0 & (i \neq j) \end{cases}$$

Example (Dual basis of the standard basis of $\mathbf{F}^n$)**.** Fix a positive integer n. For $i = 1,\dots,n$, define ϕ_i to be the linear functional on $\mathbf{F}^n$ that selects the i-th coordinate of a vector in $\mathbf{F}^n$:

$$\phi_i(x_1,\dots,x_n) = x_i$$

for each $(x_1,\dots,x_n) \in \mathbf{F}^n$.
Let $\{e_1,\dots,e_n\}$ be the standard basis of $\mathbf{F}^n$. Then

$$\phi_i(e_j) = \begin{cases} 1 & (i = j) \\ 0 & (i \neq j) \end{cases}$$

Thus $\phi_1,\dots,\phi_n$ is the dual basis of the standard basis $e_1,\dots,e_n$ of $\mathbf{F}^n$.

The next result shows that the dual basis of a basis of V consists of the linear functionals on V that give the coefficients for expressing a vector in V as a linear combination of the basis vectors.

Proposition 4.65. *Suppose $\{v_1,\dots,v_n\}$ is a basis of V, and $\{\phi_1,\dots,\phi_n\}$ is the dual basis. Then for each $v \in V$,*

$$v = \phi_1(v)v_1 + \cdots + \phi_n(v)v_n.$$

Proof. Let $v \in V$. Since $\{v_1,\dots,v_n\}$ is a basis of V, there exist $c_1,\dots,c_n \in \mathbf{F}$ such that

$$v = c_1v_1 + \cdots + c_nv_n.$$

For $i = 1,\dots,n$, applying ϕ_i to both sides of the equation above gives

$$\phi_i(v) = c_i.$$

□

The next result shows that the dual basis is a basis of the dual space. Thus the terminology "dual basis" is justified.

Lemma 4.66. *Suppose V is finite-dimensional. Then the dual basis of a basis of V is a basis of V'.*

Proof. Suppose $\{v_1,\dots,v_n\}$ is a basis of V. Let $\{\phi_1,\dots,\phi_n\}$ denote the dual basis.

Since $\{\phi_1,\dots,\phi_n\}$ has length $\dim V$, in order to show that it is a basis of V', it suffices to show that it is linearly independent in V'.

Suppose $a_1,\dots,a_n \in \mathbf{F}$ are such that

$$a_1\phi_1 + \cdots + a_n\phi_n = 0. \tag{I}$$

Now for each $i = 1,\dots,n$,

$$(a_1\phi_1 + \cdots + a_n\phi_n)(v_i) = a_i.$$

Thus (I) shows that $a_1 = \cdots = a_n = 0$. Hence $\{\phi_1,\dots,\phi_n\}$ is linearly independent. □

Definition 4.67 (Dual map)**.** Suppose $T \in \mathscr{L}(V,W)$. The ***dual map*** of T is the linear map

$$\begin{aligned} T' : W' &\to V' \\ \phi &\mapsto \phi \circ T \end{aligned}$$

If $T \in \mathscr{L}(V,W)$ and $\phi \in W'$, then $T'(\phi)$ is defined above to be the composition of the linear maps ϕ and T. Thus $T'(\phi)$ is indeed a linear map from V to $\mathbf{F}$, i.e., $T'(\phi) \in V'$.

Lemma. $T' \in \mathscr{L}(W',V')$.

Proof. Let $\phi, \psi \in W'$, $\lambda \in \mathbf{F}$.

(i) $T'(\phi+\psi) = (\phi+\psi)\circ T = \phi \circ T + \psi \circ T = T'(\phi) + T'(\psi)$.

(ii) $T'(\lambda\phi) = (\lambda\phi)\circ T = \lambda(\phi \circ T) = \lambda(T'(\phi))$.

□

Lemma 4.68 (Algebraic properties of dual map)**.** *Suppose $T \in \mathscr{L}(V,W)$. Then*

(i) $(S+T)' = S' + T'$ *for all* $S \in \mathscr{L}(V,W)$

(ii) $(\lambda T)' = \lambda T'$ *for all* $\lambda \in \mathbf{F}$

(iii) $(ST)' = T'S'$ *for all* $S \in \mathscr{L}(W,U)$

Proof.

(i)

(ii)

(iii) Let $\phi \in U'$. Then

$$(ST)'(\phi) = \phi \circ (ST) = (\phi \circ S) \circ T = T'(\phi \circ S) = T'(S'(\phi)) = (T'S')(\phi).$$

□

(i) and (ii) imply that the function that takes T to T' is a linear map from $\mathscr{L}(V,W)$ to $\mathscr{L}(W',V')$.

Kernel and Image of Dual of Linear Map

The goal of this section is to describe $\ker T'$ and $\operatorname{im} T'$ in terms of $\operatorname{im} T$ and $\ker T$. To do this, we will need the next definition.

Definition 4.69 (Annihilator). For $U \subset V$, the ***annihilator*** of U is defined by

$$U^0 := \{\phi \in V' \mid \phi(u) = \mathbf{0}, \forall u \in U\}.$$

Example. $\{\mathbf{0}\}^0 = V'$ and $V^0 = \{\mathbf{0}\}$.

We check that $U^0 \leq V$:

(i) Note that $0 \in U^0$ (here 0 is the zero linear functional on V) because the zero linear functional applied to every vector in U equals $\mathbf{0} \in \mathbf{F}$.

(ii) Suppose $\phi, \psi \in U^0$. Thus $\phi, \psi \in V'$ and $\phi(u) = \psi(u) = \mathbf{0}$ for every $u \in U$.

Let $u \in U$, then

$$(\phi + \psi)(u) = \phi(u) + \psi(u) = \mathbf{0} + \mathbf{0} = \mathbf{0}.$$

Thus $\phi + \psi \in U^0$, so U^0 is closed under addition.

(iii) Suppose $\phi \in U^0$, $\lambda \in \mathbf{F}$, let $u \in U$, then

$$\phi(\lambda u) = \lambda \phi(u) = \mathbf{0}$$

so $\lambda\phi \in U^0$, so U^0 is closed under scalar multiplication.

Lemma 4.70 (Dimension of annihilator). *Suppose V is finite-dimensional, and $U \leq V$. Then*

$$\dim U^0 = \dim V - \dim U.$$

Proof. Let $i \in \mathscr{L}(U,V)$ be the inclusion map defined by $i(u) = u$ for each $u \in U$. Thus the dual map i' is a linear map from V' to U'. The fundamental theorem of linear maps applied to i' shows that

$$\dim \ker i' + \dim \operatorname{im} i' = \dim V'. \tag{I}$$

However, $\ker i' = U^0$ (as can be seen by thinking about the definitions) and $\dim V' = \dim V$ (by 4.63), so we can rewrite (I) as

$$\dim U^0 + \dim \operatorname{im} i' = \dim V. \tag{II}$$

If $\phi \in U'$, then ϕ can be extended to a linear functional ψ on V (see, for example, Exercise 13 in Section 3A). The definition of i' shows that $i'(\psi) = \phi$. Thus $\phi \in \operatorname{im} i'$, which implies that $\operatorname{im} i' = U'$. Hence

$$\dim \ker i' = \dim U' = \dim U,$$

and then (II) becomes the equation $\dim U + \dim U^0 = \dim V$, as desired. □

The next result provides conditions for the annihilator to equal $\{\mathbf{0}\}$ or the whole space.

Lemma 4.71. *Suppose V is finite-dimensional, and $U \leq V$. Then*

(i) $U^0 = \{\mathbf{0}\} \iff U = V$

(ii) $U^0 = V' \iff U = \{\mathbf{0}\}$

Proof.

(i)

$$\begin{aligned} U^0 = \{\mathbf{0}\} &\iff \dim U^0 = 0 \\ &\iff \dim U = \dim V && \text{[by 4.70]} \\ &\iff U = V \end{aligned}$$

(ii)

$$\begin{aligned} U^0 = V' &\iff \dim U^0 = \dim V' \\ &\iff \dim U^0 = \dim V && \text{[by 4.63]} \\ &\iff \dim U = 0 && \text{[by 4.70]} \\ &\iff U = \{\mathbf{0}\} \end{aligned}$$

□

The next result concerns $\ker T'$.

Lemma 4.72. *Suppose V and W are finite-dimensional, $T \in \mathscr{L}(V,W)$. Then*

(i) $\ker T' = (\operatorname{im} T)^0$

(ii) $\dim \ker T' = \dim \ker T + \dim W - \dim V$

Proof.

(i) $\boxed{\subset}$ Let $\phi \in \ker T'$. Then $0 = T'(\phi) = \phi \circ T$. Hence

$$0 = (\phi \circ T)(v) = \phi(Tv) \quad (\forall v \in V).$$

Thus $\phi \in (\operatorname{im} T)^0$. This implies that $\ker T' \subset (\operatorname{im} T)^0$.

$\boxed{\supset}$ Let $\phi \in (\operatorname{im} T)^0$. Then $\phi(Tv) = 0$ for every $v \in V$. Hence $0 = \phi \circ T = T'(\phi)$, i.e., $\phi \in \ker T'$. Thus $\phi \in \ker T'$, which shows that $(\operatorname{im} T)^0 \subset \ker T'$.

(ii) We have

$$\begin{aligned}
\dim \ker T' &= \dim(\operatorname{im} T)^0 && \text{[by (i)]}\\
&= \dim W - \dim \operatorname{im} T && \text{[by 4.70]}\\
&= \dim W - (\dim W - \dim \ker T) && \text{[by fundamental theorem of linear maps]}\\
&= \dim \ker T + \dim W - \dim V.
\end{aligned}$$

□

The next result can be useful because sometimes it is easier to verify that T' is injective than to show directly that T is surjective.

Lemma 4.73. *Suppose V and W are finite-dimensional, $T \in \mathscr{L}(V,W)$. Then*

$$T \text{ is surjective} \iff T' \text{ is injective.}$$

Proof. Let $T \in \mathscr{L}(V,W)$. We have

$$\begin{aligned}
T \text{ is surjective} &\iff \operatorname{im} T = W\\
&\iff (\operatorname{im} T)^0 = \{\mathbf{0}\} && \text{[by 4.71]}\\
&\iff \operatorname{im} T' = \{\mathbf{0}\} && \text{[by 4.72]}\\
&\iff T' \text{ is injective}
\end{aligned}$$

□

The following result concerns $\operatorname{im} T'$.

Lemma 4.74. *Suppose V and W finite-dimensional, $T \in \mathscr{L}(V,W)$. Then*

(i) $\dim \operatorname{im} T' = \dim \operatorname{im} T$

(ii) $\operatorname{im} T' = (\ker T)^0$

Proof.

(i) We have

$$\begin{aligned}
\dim \operatorname{im} T' &= \dim W' - \dim \ker T' && \text{[by fundamental theorem of linear maps]}\\
&= \dim W - \dim(\ker T)^0 && \text{[by 4.63 and 4.72]}\\
&= \dim \operatorname{im} T && \text{[by 4.70]}
\end{aligned}$$

(ii) We first show that $\operatorname{im} T' \subset (\ker T)^0$.

Let $\phi \in \ker T'$. Then there exists $\psi \in W'$ such that $\phi = T'(\psi)$.

If $v \in \ker T$, then

$$\phi(v) = \big(T'(\psi)\big)\, v = (\psi \circ T)(v) = \psi(Tv) = \psi(\mathbf{0}) = \mathbf{0}.$$

Hence $\phi \in (\ker T)^0$. This implies that $\operatorname{im} T' \in (\ker T)^0$.

We will complete the proof by showing that $\operatorname{im} T'$ and $(\ker T)^0$ have the same dimension. To do this, note that

$$\begin{aligned} \dim \operatorname{im} T' &= \dim \operatorname{im} T && \text{[by 4.63]} \\ &= \dim V - \ker T && \text{[by fundamental theorem of linear maps]} \\ &= \dim (\ker T)^0 && \text{[by 4.70]} \end{aligned}$$

□

Lemma 4.75. *Suppose V and W are finite-dimensional, $T \in \mathscr{L}(V,W)$. Then*

$$T \text{ is injective} \iff T' \text{ is surjective.}$$

Proof. Let $T \in \mathscr{L}(V,W)$. We have

$$\begin{aligned} T \text{ is injective} &\iff \ker T = \{\mathbf{0}\} \\ &\iff (\ker T)^0 = V' && \text{[by 4.71]} \\ &\iff \operatorname{im} T' = V' && \text{[by 4.74]} \end{aligned}$$

□

Matrix of Dual of Linear Map

The setting for the next result is the assumption that we have a basis $\{v_1, \ldots, v_n\}$ of V, along with its dual basis $\{\phi_1, \ldots, \phi_n\}$ of V'. We also have a basis $\{w_1, \ldots, w_m\}$ of W, along with its dual basis $\{\psi_1, \ldots, \psi_m\}$ of W'.

Thus $\mathscr{M}(T)$ is computed with respect to the aforementioned bases of V and W, and $\mathscr{M}(T')$ is computed with respect to the aforementioned dual bases of W' and V'. Using these bases gives the following result.

Lemma 4.76. *Suppose V and W are finite-dimensional, $T \in \mathscr{L}(V,W)$. Then*

$$\mathscr{M}(T') = \mathscr{M}(T)^{\top}.$$

Proof. Let $\mathscr{M}(T) = A$, $\mathscr{M}(T') = C$. From the definition of $\mathscr{M}(T')$ we have

$$T'(\psi_i) = \sum_{k=1}^{n} C_{ki}\phi_k.$$

The left side of the equation above equals $\psi_i \circ T$. Thus applying both sides of the equation above to v_j gives

$$\begin{aligned}(\psi_i \circ T)(v_j) &= \sum_{k=1}^{n} C_{ki}\phi_k(v_j) \\ &= C_{ji}.\end{aligned}$$

We also have

$$\begin{aligned}(\psi_i \circ T)(v_j) &= \psi_i(Tv_j) \\ &= \psi_i\left(\sum_{k=1}^{m} A_{kj}w_k\right) \\ &= \sum_{k=1}^{m} A_{kj}\psi_i(w_k) \\ &= A_{ij}.\end{aligned}$$

Comparing the last line of the last two sets of equations, we have $C_{j,i} = A_{i,j}$. Thus $C = A^\top$, so $\mathscr{M}(T') = \mathscr{M}(T)^\top$ as desired. □

Exercises

Exercise 4.1 ([Axl24] 3A)**.** Suppose $b, c \in \mathbb{R}$. Define $T\colon \mathbb{R}^3 \to \mathbb{R}^2$ by

$$T(x,y,z) = (2x-4y+3z+b, 6x+cxyz).$$

Show that T is linear if and only if $b = c = 0$.

Exercise 4.2 ([Axl24] 3A Q11)**.** Suppose V is finite-dimensional, $T \in \mathscr{L}(V)$. Prove that T is a scalar multiple of the identity if and only if $ST = TS$ for all $S \in \mathscr{L}(V)$.

Exercise 4.3 ([Axl24] 3B Q9)**.** Suppose $T \in \mathscr{L}(V,W)$ is injective, $\{v_1, \dots, v_n\}$ is linearly independent in V. Prove that $\{Tv_1, \dots, Tv_n\}$ is linearly independent in W.

Solution. Suppose there exist $a_i \in \mathbf{F}$ such that

$$\begin{aligned} & a_1 Tv_1 + \cdots + a_n Tv_n = \mathbf{0} \\ \implies & T(a_1 v_1 + \cdots + a_n v_n) = 0 \\ \implies & a_1 v_1 + \cdots + a_n v_n \in \ker T \end{aligned}$$

Since T is injective,

$$\ker T = \{\mathbf{0}\} \implies a_1 v_1 + \cdots + a_n v_n = \mathbf{0} \implies a_1 = \cdots = a_n = 0$$

since $\{v_1, \dots, v_n\}$ is linearly independent. □

Exercise 4.4 ([Axl24] 3B Q11)**.** Suppose that V is finite-dimensional, $T \in \mathscr{L}(V,W)$. Prove that there exists $U \leq V$ such that

$$U \cap \ker T = \{\mathbf{0}\} \quad \text{and} \quad \operatorname{im} T = T(U).$$

Solution. □

Exercise 4.5 ([Axl24] 3B Q19)**.** Suppose W is finite-dimensional, $T \in \mathscr{L}(V,W)$. Prove that T is injective if and only if there exists $S \in \mathscr{L}(W,V)$ such that ST is the identity operator on V.

Solution. □

Exercise 4.6 ([Axl24] 3B Q20)**.** Suppose W is finite-dimensional, $T \in \mathscr{L}(V,W)$. Prove that T is surjective if and only if there exists $S \in \mathscr{L}(W,V)$ such that TS is the identity operator on W.

Exercise 4.7 ([Axl24] 3B 22)**.** Suppose U,V are finite-dimensional, $S \in \mathscr{L}(V,W)$, $T \in \mathscr{L}(U,V)$. Prove that

$$\dim\ker ST \leq \dim\ker S + \dim\ker T.$$

Solution. □

Exercise 4.8 ([Axl24] 3D)**.** Suppose $T \in \mathscr{L}(V,W)$ is invertible. Show that T^{-1} is invertible and

$$\left(T^{-1}\right)^{-1} = T.$$

Solution. T^{-1} is invertible because there exists T such that $TT^{-1} = T^{-1}T = I$. So

$$T^{-1}T = TT^{-1} = I$$

thus $\left(T^{-1}\right)^{-1} = T$. □

3C Q15,16,17

3D Q11,12,17,22,23,24

Exercise 4.9 ([Axl24] 3D)**.** Suppose $T \in \mathscr{L}(U,V)$ and $S \in \mathscr{L}(V,W)$ are both invertible linear maps. Prove that $ST \in \mathscr{L}(U,W)$ is invertible and that $(ST)^{-1} = T^{-1}S^{-1}$.

Solution.

$$(ST)(T^{-1}S^{-1}) = S(TT^{-1})S^{-1} = I = T^{-1}S^{-1}ST.$$

□

Exercise 4.10 ([Axl24] 3D)**.** Suppose V is finite-dimensional and $T \in \mathscr{L}(V,W)$. Prove that the following are equivalent:

(i) T is invertible;

(ii) $\{Tv_1,\ldots,Tv_n\}$ is a basis of V for every basis $\{v_1,\ldots,v_n\}$ of V;

(iii) $\{Tv_1,\ldots,Tv_n\}$ is a basis of V for some basis $\{v_1,\ldots,v_n\}$ of V.

Solution.

(i) $\implies$ (ii) It only suffices to prove linear independence. We can show this

$$a_1Tv_1 + \cdots + a_nTv_n = 0 \iff a_1v_1 + \cdots + a_nv_n = 0$$

since T is injective and thus the only solution is all a_i are identically zero.

(ii) $\implies$ (iii) Trivial.

(iii) $\implies$ (i) By the linear map lemma, there exists $S \in \mathscr{L}(V)$ such that $S(Tv_i) = v_i$ for all i. Such S is the inverse of T (one can verify) and thus T is invertible. □

Exercise 4.11 ([Axl24] 3E Q3)**.** Suppose $V_1, \ldots, V_m$ are vector spaces. Prove that

$$\mathscr{L}(V_1 \times \cdots \times V_m, W) \cong \mathscr{L}(V_1, W) \times \cdots \times \mathscr{L}(V_m, W).$$

Exercise 4.12 ([Axl24] 3E Q4)**.** Suppose $V_1, \ldots, V_m$ are vector spaces. Prove that

$$\mathscr{L}(V, W_1 \times \cdots \times W_m) \cong \mathscr{L}(V, W_1) \times \cdots \times \mathscr{L}(V, W_m).$$

Exercise 4.13 ([Axl24] 3E Q5)**.** For a positive integer m, define V^m by

$$V^m = \underbrace{V \times \cdots \times V}_{m \text{ times}}.$$

Prove that $V^m \cong \mathscr{L}(\mathbf{F}^m, V)$.

Exercise 4.14 ([Axl24] 3E Q6)**.** Suppose that $v, x \in V$ and $U, W \leq V$ are such that $v + U = x + W$. Prove that $U = W$.

Exercise 4.15 ([Axl24] 3E Q12, Barycentric coordinates)**.** Suppose $v_1, \ldots, v_m \in V$. Let

$$A = \{\lambda_1 v_1 + \cdots + \lambda_m v_m \mid \lambda_i \in \mathbf{F}, \lambda_1 + \cdots + \lambda_m = 1\}.$$

(i) Prove that A is a coset of some subspace of V.

(ii) Prove that if B is a coset of some subspace of V, and $\{v_1, \ldots, v_m\} \subset B$, then $A \subset B$.

(iii) Prove that A is a coset of some subspace of V, where $\dim V < m$.

Exercise 4.16 ([Axl24] 3E Q13)**.** Suppose $U \leq V$, and V/U is finite-dimensional. Prove that $V \cong U \times (V/U)$.

Solution.

$$\dim V = \dim U + (\dim V - \dim U) = \dim U + \dim(V/U).$$

□

Exercise 4.17 ([Axl24] 3E Q14)**.** Suppose $U, W \leq V$ such that $V = U \oplus W$. Suppose $w_1, \ldots, w_m$ is a basis of W. Prove that $w_1 + U, \ldots, w_m + U$ is a basis of V/U.

Exercise 4.18 ([Axl24] 3E Q15)**.**

Exercise 4.19 ([Axl24] 3E Q16)**.** Suppose $\phi \in \mathscr{L}(V, \mathbf{F})$ and $\phi \neq 0$. Prove that $\dim V / \ker \phi = 1$.

Exercise 4.20 ([Axl24] 3E Q18)**.**

Exercise 4.21 ([Axl24] 3E Q19)**.** Suppose $T \in \mathscr{L}(V, W)$ and $U \leq V$. Let π denote the quotient map from V to V/U. Prove that there exists $S \in \mathscr{L}(V/U, W)$ such that

$$T = S \circ \pi \iff U \subset \ker T.$$

5 Polynomials

5.1 Definitions

Definition 5.1 (Polynomial)**.** We say $p : \mathbf{F} \to \mathbf{F}$ is a ***polynomial*** with coefficients in $\mathbf{F}$ if there exist $a_i \in \mathbf{F}$ such that

$$p(z) = a_0 + a_1 z + \cdots + a_n z^n \quad (z \in \mathbf{F})$$

Notation. Let $\mathbf{F}[z]$ denote the set of polynomials with coefficients in $\mathbf{F}$.

Lemma 5.2. *With the usual operations of addition and scalar multiplication, $\mathbf{F}[z]$ is a vector space over $\mathbf{F}$.*

Hence $\mathbf{F}[z]$ is a subspace of $\mathbf{F}^{\mathbf{F}}$ (vector space of functions from $\mathbf{F}$ to $\mathbf{F}$).

Definition 5.3 (Degree)**.** A polynomial $p \in \mathbf{F}[z]$ is has ***degree*** n, denoted by $\deg p = n$, if there exist scalars $a_0, a_1, \ldots, a_n \in \mathbf{F}$ with $a_n \neq 0$ such that $p(z) = a_0 + a_1 z + \cdots + a_n z^n$ for all $z \in \mathbf{F}$.

Notation. For non-negative integer n, $\mathbf{F}_n[z]$ denotes the set of polynomials with coefficients in $\mathbf{F}$ and degree at most n.

Lemma 5.4. *For non-negative integer n, $\mathbf{F}_n[z]$ is finite-dimensional.*

Proof. $\mathbf{F}_n[z] = \operatorname{span}(1, z, z^2, \ldots, z^n)$ [here we slightly abuse notation by letting z^k denote a function]. □

Lemma 5.5. *$\mathbf{F}[z]$ is infinite-dimensional.*

Proof. Consider any list of elements of $\mathbf{F}[z]$. Let n denote the highest degree of the polynomials in this list. Then every polynomial in the span of this list has degree at most n. Thus z^{n+1} is not in the span of our list. Hence no list spans $\mathbf{F}[z]$. Thus $\mathbf{F}[z]$ is infnite-dimensional. □

5.2 Zeros of Polynomials

Definition 5.6 (Zero of polynomial)**.** We call $\lambda \in \mathbf{F}$ a ***zero*** of a polynomial $p \in \mathbf{F}[z]$ if

$$p(\lambda) = 0.$$

Lemma 5.7 (Factor theorem)**.** *Suppose $n \in \mathbb{N}$, $p \in \mathbf{F}_n[z]$. Suppose $\lambda \in \mathbf{F}$, then $p(\lambda) = 0$ if and only if there exists $q \in \mathbf{F}_{n-1}[z]$ such that*

$$p(z) = (z-\lambda)q(z) \quad (z \in \mathbf{F}).$$

Proof.

$\boxed{\Longrightarrow}$ Suppose $p(\lambda) = 0$. Let $a_0, a_1, \ldots, a_n \in \mathbf{F}$ be such that

$$p(z) = a_n z^n + \cdots + a_1 z + a_0 \quad (z \in \mathbf{F}).$$

Then for all $z \in \mathbf{F}$,

$$\begin{aligned} p(z) &= p(z) - p(\lambda) \\ &= (a_n z^n + \cdots + a_1 z + a_0) - (a_n \lambda^n + \cdots + a_1 \lambda + a_0) \\ &= a_n (z^n - \lambda^n) + \cdots + a_1 (z - \lambda). \end{aligned}$$

Note that for each $k = 1, \ldots, n$, we can factorise

$$z^k - \lambda^k = (z-\lambda)\left(z^{k-1} + z^{k-2}\lambda + \cdots + \lambda^{k-1}\right).$$

Thus p equals $z - \lambda$ times some polynomial of degree $n-1$, as desired.

$\boxed{\Longleftarrow}$ Suppose there exists $q \in \mathbf{F}[z]$ such that

$$p(z) = (z-\lambda)q(z) \quad (z \in \mathbf{F}).$$

Then

$$p(\lambda) = (\lambda - \lambda)q(\lambda) = 0,$$

as desired. □

Now we can prove that a polynomial has at most as many zeros as its degree.

Proposition 5.8. *Suppose $n \in \mathbb{N}$, $p \in \mathbf{F}_n[z]$. Then p has at most n zeros in $\mathbf{F}$.*

Proof. Prove by induction on n.

The desired result holds for $n = 1$ because if $a_1 \neq 0$ then the polynomial $a_0 + a_1 z$ has only one zero (which equals $-\frac{a_0}{a_1}$).

Now assume the desired result holds for $n-1$. If p has no zeros in $\mathbf{F}$, then the desired result holds and we are done. Thus suppose p has a zero $\lambda \in \mathbf{F}$. By 5.7, there exists $q \in \mathbf{F}[z]$ of degree $n-1$ such that

$$p(z) = (z-\lambda)q(z) \quad (\forall z \in \mathbf{F})$$

By the induction hypothesis, q has at most $n-1$ zeros in $\mathbf{F}$. The equation above shows that the zeros of p in $\mathbf{F}$ are exactly the zeros of q in $\mathbf{F}$ along with λ. Thus p has at most n zeros in $\mathbf{F}$. □

The result above implies that the coefficients of a polynomial are uniquely determined (because if a polynomial had two different sets of coefficients, then subtracting the two representations of the polynomial would give a polynomial with some nonzero coefficients but infinitely many zeros). In particular, the degree of a polynomial is uniquely defined.

5.3 Division Algorithm for Polynomials

Proposition 5.9 (Division algorithm). *Suppose $p, s \in \mathbf{F}[z]$, $s \neq 0$. Then there exists unique polynomials $q, r \in \mathbf{F}[z]$, where $\deg r < \deg s$, such that*

$$p = sq + r.$$

Proof. Let $n = \deg p$, $m = \deg s$. If $n < m$, take $q = 0$ and $r = p$ to get the desired equation.

Now assume that $n \geq m$. The set

$$S = \{1, z, \ldots, z^{m-1}, s, zs, \ldots, z^{n-m}s\}$$

is linearly independent in $\mathbf{F}[z]$ because each polynomial in S has a different degree. Also, S has length $n+1$, which equals $\dim \mathbf{F}[z]$. Hence S is a basis of $\mathbf{F}[z]$.

Since $p \in \mathbf{F}[z]$ and S is a basis of $\mathbf{F}[z]$, there exist unique constants $a_0, a_1, \ldots, a_{m-1} \in \mathbf{F}$ and $b_0, b_1, \ldots, b_{n-m} \in \mathbf{F}$ such that

$$\begin{aligned} p &= a_0 + a_1 z + \cdots + a_{m-1} z^{m-1} + b_0 s + b_1 z s + \cdots + b_{n-m} z^{n-m} s \\ &= \underbrace{a_0 + a_1 z + \cdots + a_{m-1} z^{m-1}}_{r} + s \left(\underbrace{b_0 + b_1 z + \cdots + b_{n-m} z^{n-m}}_{q} \right). \end{aligned}$$

With r and q as defined above, we see that p can be written as $p = sq + r$ with $\deg r < \deg s$, as desired.

The uniqueness of $q, r \in \mathbf{F}[z]$ satisfying these conditions follows from the uniqueness of the constants $a_0, a_1, \ldots, a_{m-1} \in \mathbf{F}$ and $b_0, b_1, \ldots, b_{n-m} \in \mathbf{F}$. □

5.4 Factorisation of Polynomials over $\mathbb{C}$

Theorem 5.10 (Fundamental theorem of algebra I)**.** *Every non-constant polynomial with complex coefficients has a zero in* $\mathbb{C}$.

Proof. Recall de Moivre's theorem: if $k \in \mathbb{Z}^+$ and $\theta \in \mathbb{R}$, then

$$(\cos\theta + i\sin\theta)^k = \cos k\theta + i\sin k\theta.$$

Let $w \in \mathbb{C}$, and $k \in \mathbb{Z}^+$. We can express w in polar coordinates:

$$r(\cos\theta + i\sin\theta) = w$$

for some $r \in \mathbb{R}^+$, $\theta \in \mathbb{R}$. Applying de Moivre's theorem to the LHS gives

$$\left[r^{\frac{1}{k}}\left(\cos\frac{\theta}{k} + i\sin\frac{\theta}{k}\right)\right]^k = w.$$

Thus every complex number has a k-th root, a fact that we will soon use.

Suppose p is a non-constant polynomial with complex coefficients and highest-order non-zero term $c_m z^m$. Then $|p(z)| \to \infty$ as $|z| \to \infty$ (because $|p(z)|/|z^m| \to |c_m|$ as $|z| \to \infty$). Thus the continuous function $z \mapsto |p(z)|$ has a global minimum at some point $\xi \in \mathbb{C}$.

Claim. $p(\xi) = 0$.

Suppose, for a contradiction, that $p(\xi) \neq 0$. Define a new polynomial q by

$$q(z) = \frac{p(z+\xi)}{p(\xi)}.$$

The function $z \mapsto |q(z)|$ has a global minimum of 1 at $z = 0$. Write

$$q(z) = 1 + a_k z^k + \cdots + a_m z^m,$$

where k is the smallest positive integer such that the coefficient of z^k is non-zero; in other words, $a_k \neq 0$.

Let $\beta \in \mathbb{C}$ be such that $\beta^k = -\dfrac{1}{a_k}$. There is a constant $c > 1$ such that if $t \in (0,1)$, then

$$\begin{aligned} |q(t\beta)| &\leq |1 + a_k t^k \beta^k| + t^{k+1} c \\ &= 1 - t^k(1 - tc). \end{aligned}$$

Taking $t = \frac{1}{2c}$, we have $|q(t\beta)| < 1$, which contradicts the assumption that the global minimum of $z \mapsto |q(z)|$ is 1. Hence $p(\xi) = 0$, which shows that p has a zero. □

Remark. The fundamental theorem of algebra is an existence theorem. Its proof does not lead to a method for finding zeros.

The first version of the fundamental theorem of algebra leads to the following factorisation result for polynomials with complex coefficients. Note that in this factorisation, the zeros of p are the numbers $\lambda_1, \ldots, \lambda_n$, which are the only values of z for which the right side of the equation in the next result equals 0.

Theorem 5.11 (Fundamental theorem of algebra II). *If $p \in \mathbb{C}[z]$ is a non-constant polynomial, then p has a unique factorisation (except for the order of the factors) of the form*

$$p(z) = c(z-\lambda_1)\cdots(z-\lambda_n),$$

where $c, \lambda_1, \ldots, \lambda_n \in \mathbb{C}$.

Proof. Let $p \in \mathbb{C}[z]$, and let $n = \deg p$. We shall induct on n.

If $n = 1$, then the desired factorisation exists and is unique.

Suppose $n > 1$ and the desired factorisation exists and is unique for all polynomials of degree $n-1$.

Existence By the first version of the fundamental theorem of algebra (5.10), p has a zero $\lambda \in \mathbb{C}$. By the factor theorem, there is a polynomial q of degree $n-1$ such that

$$p(z) = (z-\lambda)q(z) \quad (z \in \mathbb{C}).$$

By induction hypothesis, q has the desired factorisation, which when plugged into the equation above gives the desired factorisation of p.

Uniqueness The number c is uniquely determined as the coefficient of z^n in p. Thus we only need to show that except for the order, there is only one way to choose $\lambda_1, \ldots, \lambda_n$. Suppose

$$(z-\lambda_1)\cdots(z-\lambda_n) = (z-\tau_1)\cdots(z-\tau_n) \quad (z \in \mathbb{C}).$$

When $z = \lambda_1$, the LHS equals 0, so the RHS equals 0. Thus one of the τ's on the RHS must equal λ_1. Relabelling, we can assume $\tau_1 = \lambda_1$. Now if $z \neq \lambda_1$, we can divide both sides of the equation by $z-\lambda_1$ to get

$$(z-\lambda_2)\cdots(z-\lambda_n) = (z-\tau_2)\cdots(z-\tau_n)$$

for all $z \in \mathbb{C}$ except possibly $z = \lambda_1$. Actually the equation above holds for all $z \in \mathbb{C}$, because otherwise by subtracting the right side from the left side we would get a non-zero polynomial that has infinitely many zeros.

The equation above and our induction hypothesis imply that except for the order, the λ's are the same as the τ's, completing the proof of uniqueness. □

5.5 Factorisation of Polynomials over $\mathbb{R}$

A polynomial with real coefficients may have no real zeros. For example, the polynomial x^2+1 has no real zeros.

To obtain a factorisation theorem over $\mathbb{R}$, we will use our factorisation theorem over $\mathbb{C}$. We begin with the next result: polynomials with real coefficients have non-real zeros in pairs.

Lemma 5.12. *Suppose $p \in \mathbb{C}[z]$ is a polynomial with real coefficients. If $\lambda \in \mathbb{C}$ is a zero of p, then so is its conjugate $\overline{\lambda}$.*

Proof. Let

$$p(z) = a_0 + a_1 z + \cdots + a_n z^n,$$

where $a_0, \ldots, a_n \in \mathbb{R}$. Suppose $\lambda \in \mathbb{C}$ is a zero of p, then

$$a_0 + a_1\lambda + \cdots + a_n\lambda^n = 0.$$

Taking the complex conjugate on both sides of the equation gives

$$a_0 + a_1\overline{\lambda} + \cdots + a_n\overline{\lambda}^n = 0.$$

Hence $\overline{\lambda}$ is a zero of p. □

We want a factorisation theorem for polynomials with real coefficients. We begin with the following result.

Remark. Think about the quadratic formula in connection with the result below.

Lemma 5.13 (Factorisation of quadratic polynomial)**.** *Suppose $b, c \in \mathbb{R}$. Then there is a polynomial factorisation of the form*

$$x^2 + bx + c = (x - \lambda_1)(x - \lambda_2)$$

with $\lambda_1, \lambda_2 \in \mathbb{R}$ if and only if $b^2 \geq 4c$.

Proof. Completing the square gives

$$x^2 + bx + c = \left(x + \frac{b}{2}\right)^2 + \left(c - \frac{b^2}{4}\right). \tag{I}$$

$\boxed{\Longrightarrow}$ We prove the contrapositive. Suppose $b^2 < 4c$, then the RHS of (I) is positive for every $x \in \mathbb{R}$. Hence the polynomial $x^2 + bx + c$ has no real zeros and thus cannot be factored in the form $(x - \lambda_1)(x - \lambda_2)$ with $\lambda_1, \lambda_2 \in \mathbb{R}$.

$\boxed{\Longleftarrow}$ Suppose $b^2 \geq 4c$. Then there is a real number d such that $d^2 = \frac{b^2}{4} - c$. We can rewrite (I) as

$$\begin{aligned} x^2 + bx + c &= \left(x + \frac{b}{2}\right)^2 - d^2 \\ &= \left(x + \frac{b}{2} + d\right)\left(x + \frac{b}{2} - d\right), \end{aligned}$$

which gives the desired factorisation. □

Theorem 5.14 (Factorisation of polynomial over $\mathbb{R}$). *Suppose $p \in \mathbb{R}[x]$ is a non-constant polynomial. Then p has a unique factorisation (except for the order of the factors) of the form*

$$p(x) = c(x-\lambda_1)\cdots(x-\lambda_n)(x^2+b_1x+c_1)\cdots(x^2+b_Nx+c_N),$$

where $c, \lambda_1, \ldots, \lambda_n, b_1, \ldots, b_N, c_1, \ldots, c_N \in \mathbb{R}$, with ${b_i}^2 < 4c_i$ for each i.

Proof.

Existence Consider p as an element of $\mathbb{C}[x]$. If all (complex) zeros of p are real, then we have the desired factorisation by 5.11. Thus suppose p has a zero $\lambda \in \mathbb{C}$ with $\lambda \notin \mathbb{R}$. By 5.12, $\overline{\lambda}$ is a zero of p. Thus we can write

$$\begin{aligned} p(x) &= (x-\lambda)(x-\overline{\lambda})q(x) \\ &= (x^2 - 2(\operatorname{Re}\lambda)x + |\lambda|^2)q(x) \end{aligned}$$

for some polynomial $q \in \mathbb{C}[x]$ of degree two less than the degree of p. If we can prove that q has real coefficients, then using induction on the degree of p completes the proof of the existence part of this result.

To prove that q has real coefficients, we solve the equation above for q, getting

$$q(x) = \frac{p(x)}{x^2 - 2(\operatorname{Re}\lambda)x + |\lambda|^2} \quad (x \in \mathbb{R}).$$

The equation above implies that $q(x) \in \mathbb{R}$ for all $x \in \mathbb{R}$. Writing

$$q(x) = a_0 + a_1x + \cdots + a_{n-2}x^{n-2},$$

where $n = \deg p$ and $a_0, \ldots, a_{n-2} \in \mathbb{C}$, we thus have

$$0 = \operatorname{Im} q(x) = (\operatorname{Im} a_0) + (\operatorname{Im} a_1)x + \cdots + (\operatorname{Im} a_{n-2})x^{n-2} \quad (x \in \mathbb{R}).$$

This implies that $\operatorname{Im} a_0, \ldots, \operatorname{Im} a_{n-2}$ all equal 0 (by 5.8). Thus all coefficients of q are real, as desired. Hence the desired factorisation exists.

Uniqueness A factor of p of the form $x^2 + b_i x + c_i$ with $b_i^2 < 4c_i$ can be uniquely written as $(x - \lambda_i)(x - \overline{\lambda_i})$ with $\lambda_i \in \mathbb{C}$. A moment's thought shows that two different factorisations of p as an element of $\mathbb{R}[x]$ would lead to two different factorisations of p as an element of $\mathbb{C}[x]$, contradicting 5.11. □

6 Eigenvalues and Eigenvectors

6.1 Invariant Subspaces

Eigenvalues

Definition 6.1 (Operator)**.** An ***operator*** is a linear map from a vector space to itself.

Definition 6.2 (Invariant subspace)**.** Suppose $T \in \mathscr{L}(V)$. We say $U \leq V$ is ***invariant*** under T if $Tu \in U$ for all $u \in U$.

Example. Suppose $T \in \mathscr{L}(V)$. Then the following subspaces of V are all invariant under T.

(i) The subspace $\{\mathbf{0}\}$ is invariant under T: if $u \in \{\mathbf{0}\}$, then $u = \mathbf{0}$ so $Tu = \mathbf{0} \in \{\mathbf{0}\}$.

(ii) The subspace V is invariant under T: if $u \in V$, then $Tu \in V$.

(iii) The subspace $\ker T$ is invariant under T: if $u \in \ker T$, then $Tu = \mathbf{0}$, and hence $Tu \in \ker T$, since a subspace must contain $\mathbf{0}$.

(iv) The subspace $\operatorname{im} T$ is invariant under T: if $u \in \operatorname{im} T$, then $Tu \in \operatorname{im} T$ by definition.

Definition 6.3 (Eigenvalue and eigenvector)**.** Suppose $T \in \mathscr{L}(V)$. $\lambda \in \mathbf{F}$ is an ***eigenvalue*** of T if there exists $v \in V \setminus \{\mathbf{0}\}$ such that $Tv = \lambda v$; we say v is an ***eigenvector*** of T corresponding to λ.

Lemma 6.4 (Equivalent conditions to be an eigenvalue)**.** *Suppose V is finite-dimensional, $T \in \mathscr{L}(V)$, $\lambda \in \mathbf{F}$. Then the following are equivalent:*

(i) λ is an eigenvalue of T.

(ii) $T - \lambda I$ *is not injective.*

(iii) $T - \lambda I$ *is not surjective.*

(iv) $T - \lambda I$ *is not invertible.*

Proof.

(i) $\iff$ (ii) Suppse λ is an eigenvalue of T, corresponding to eigenvector v. Then

$$\begin{aligned} Tv = \lambda v &\iff (T - \lambda I)v = \mathbf{0} \\ &\iff v \in \ker(T - \lambda I) \\ &\iff \ker(T - \lambda I) \neq \{\mathbf{0}\} \\ &\iff T - \lambda I \text{ is not injective.} \end{aligned}$$

(ii) $\iff$ (iii) $\iff$ (iv) This follows from 4.31. □

Proposition 6.5 (Linearly independent eigenvectors)**.** *Suppose* $T \in \mathscr{L}(V)$*. Then every set of eigenvectors of* T *corresponding to distinct eigenvalues of* T *is linearly independent.*

Proof. Let $v_1, \ldots, v_n$ be eigenvectors of T corresponding to distinct eigenvalues $\lambda_1, \ldots, \lambda_n$.

Suppose, for a contradiction, that the desired result is false. Since $v_1, \ldots, v_n$ are linearly dependent, by the linear dependence lemma, there exists a smallest positive integer $m \leq n$ such that $v_1, \ldots, v_m$ are linearly dependent.

Then there exists $a_1, \ldots, a_m \in \mathbf{F}$, none of which are 0 (because of the minimality of m) such that

$$a_1 v_1 + \cdots + a_m v_m = \mathbf{0}.$$

Applying $T - \lambda_m I$ to both sides,

$$\begin{aligned} a_1(T - \lambda_m I)v_1 + \cdots + a_{m-1}(T - \lambda_m I)v_{m-1} + a_m(T - \lambda_m I)v_m &= \mathbf{0} \\ a_1(Tv_1 - \lambda_m v_1) + \cdots + a_{m-1}(Tv_{m-1} - \lambda_m v_{m-1}) + a_m(Tv_m - \lambda_m v_m) &= \mathbf{0} \\ a_1(\lambda_1 - \lambda_m)v_1 + \cdots + a_{m-1}(\lambda_{m-1} - \lambda_m)v_{m-1} &= \mathbf{0} \end{aligned}$$

Since the eigenvalues $\lambda_1, \ldots, \lambda_m$ are distinct, none of the coefficients $a_i(\lambda_i - \lambda_m)$ equal 0. Thus $v_1, \ldots, v_{m-1}$ are $m-1$ linearly dependent eigenvectors of T corresponding to distinct eigenvalues, contradicting the minimality of m. □

An immediate corollary establishes an upper bound on the number of distinct eigenvalues that an operator can have.

Corollary 6.6. *Suppose V is finite-dimensional. Then each operator on V has at most $\dim V$ distinct eigenvalues.*

Proof. Suppose $T \in \mathscr{L}(V)$. Let $\lambda_1, \ldots, \lambda_m$ be distinct eigenvalues of T corresponding to eigenvectors $v_1, \ldots, v_m$.

By 6.5, the eigenvectors $v_1, \ldots, v_m$ are linearly independent. Since the length of a linearly independent set is less than or equal to the length of a spanning set, we have that $m \leq \dim V$, as desired. □

Polynomials Applied to Operators

Suppose $T \in \mathscr{L}(V)$. We now define powers of an operator.

- Let $n \in \mathbb{Z}^+$. Define $T^n = \underbrace{T \cdots T}_{n \text{ times}}$. Define T^0 to be the identity operator I on V.
- If T is invertible with inverse T^{-1}, define $T^{-n} = \left(T^{-1}\right)^n$.

You should verify that if T is an operator, then

$$T^m T^n = T^{m+n}, \quad (T^m)^n = T^{mn}.$$

Having defined powers of an operator, we can define what it means to apply a polynomial to an operator.

Definition 6.7. Suppose $T \in \mathscr{L}(V)$, let $p \in \mathbf{F}[z]$ be a polynomial given by

$$p(z) = a_n z^n + \cdots + a_1 z + a_0 \quad (z \in \mathbf{F})$$

Then $p(T)$ is the operator on V defined by

$$p(T) := a_n T^n + \cdots + a_1 T + a_0.$$

If we fix an operator $T \in \mathscr{L}(V)$, then the function $\mathbf{F}[z] \to \mathscr{L}(V)$ given by $p \mapsto p(T)$ is linear:

(i) $f(p+q) = (p+q)(T) = p(T) + q(T) = f(p) + f(q)$.

(ii) $f(\lambda p) = (\lambda p)(T) = \lambda p(T) = \lambda f(p)$.

Definition 6.8 (Product of polynomials). Suppose $p, q \in \mathbf{F}[z]$. Then $pq \in \mathbf{F}[z]$ is the polynomial defined by

$$(pq)(z) = p(z)q(z) \quad (z \in \mathbf{F})$$

Lemma 6.9. *Suppose $p, q \in \mathbf{F}[z]$, $T \in \mathscr{L}(V)$. Then*

(i) $(pq)(T) = p(T)q(T)$; *(multiplicativity)*

(ii) $p(T)q(T) = q(T)p(T)$. *(commutativity)*

This means when a product of polynomials is expanded using the distributive property, it does not matter whether the symbol is z or T.

Proof.

(i) Let $p(z) = \sum_{i=0}^{m} a_i z^i$, $q(z) = \sum_{j=0}^{n} b_j z^j$. Then

$$(pq)(z) = p(z)q(z) = \left(\sum_{i=0}^{m} a_i z^i\right)\left(\sum_{j=0}^{n} b_j z^j\right) = \sum_{i=0}^{m}\sum_{j=0}^{n} a_i b_j z^{i+j}.$$

Thus

$$(pq)(T) = \sum_{i=0}^{m}\sum_{j=0}^{n} a_i b_j T^{i+j} = \left(\sum_{i=0}^{m} a_i T^i\right)\left(\sum_{j=0}^{n} b_j T^j\right) = p(T)q(T).$$

(ii) Using (i) twice, we have

$$p(T)q(T) = (pq)(T) = (qp)(T) = q(T)p(T)$$

since the multiplication of polynomials is commutative.

□

Lemma 6.10. *Suppose $T \in \mathscr{L}(V)$, $p \in \mathbf{F}[z]$. Then*

(i) $\ker p(T)$ *is invariant under* T*;*

(ii) $\operatorname{im} p(T)$ *is invariant under* T.

Proof.

(i) Let $u \in \ker p(T)$. Then $p(T)u = \mathbf{0}$, so

$$(p(T))(Tu) = (p(T)T)(u) = (Tp(T))(u) = T(p(T)u) = T(\mathbf{0}) = \mathbf{0} \implies Tu \in \ker p(T).$$

Hence $\ker p(T)$ is invariant under T.

(ii) Let $u \in \operatorname{im} p(T)$. Then there exists $v \in V$ such that $u = p(T)v$. Thus

$$Tu = T(p(T)v) = p(T)(Tv) \implies Tu \in \operatorname{im} p(T).$$

Hence $\operatorname{im} p(T)$ is invariant under T.

□

6.2 The Minimal Polynomial

Existence of Eigenvalues on Complex Vector Spaces

The following is one of the most important results in linear algebra.

Theorem 6.11 (Existence of eigenvalues). *Every operator on a finite-dimensional, non-zero, complex vector space has an eigenvalue.*

Proof. Suppose V is a finite-dimensional complex vector space, $\dim V = n > 0$, $T \in \mathscr{L}(V)$. Let $v \in V \setminus \{\mathbf{0}\}$. Consider the set

$$S = \{v, Tv, T^2v, \ldots, T^n v\}.$$

Since $\dim V = n$ and S has length $n+1$, S is not linearly independent. Thus there exist $a_0, \ldots, a_n \in \mathbb{C}$, not all 0, such that

$$a_0 v + a_1 Tv + a_2 T^2 v + \cdots + a_n T^n v = \mathbf{0},$$

which we can write as

$$p(T)v = \mathbf{0},$$

where we pick p such that $\deg p$ is minimal.

By the fundamental theorem of algebra (5.10), p has a zero $\lambda \in \mathbb{C}$. By the factor theorem,

$$p(z) = (z - \lambda)q(z) \quad (z \in \mathbb{C}).$$

Thus

$$\begin{aligned} p(T) &= (T - \lambda I)q(T) \\ \mathbf{0} = p(T)v &= (T - \lambda I)q(T)v \\ Tq(T)v &= \lambda q(T)v \end{aligned}$$

Since $\deg q < \deg p$, by minimality of $\deg p$, we must have $q(T)v \neq \mathbf{0}$. Hence λ is an eigenvalue of T, with corresponding eigenvector $q(T)v$. □

Example. Note that the hypothesis in 6.11 that $\mathbf{F} = \mathbb{C}$ cannot be replaced with the hypothesis that $\mathbf{F} = \mathbb{R}$. For instance, consider $T \in \mathscr{L}(\mathbb{R}^2)$ defined by

$$Tv = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} v. \qquad (*)$$

Then

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

Notice that T is a rotation, so there is no vector that is fixed in its original direction. Hence T does not have an eigenvalue.
In contrast, consider $T \in \mathscr{L}(\mathbb{C}^2)$ defined by $(*)$. Then

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} i \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ i \end{pmatrix} = i\begin{pmatrix} i \\ 1 \end{pmatrix},$$

so i is an eigenvalue with corresponding eigenvector $\begin{pmatrix} i \\ 1 \end{pmatrix}$.

Eigenvalues and the Minimal Polynomial

A *monic polynomial* is a polynomial whose highest-degree coefficient equals 1.

The following result shows the existence, uniqueness and degree of the *minimal polynomial*.

Theorem 6.12. *Suppose V is finite-dimensional, $T \in \mathscr{L}(V)$. Then there exists a unique monic polynomial $p \in \mathbf{F}[z]$ of smallest degree such that $p(T) = 0$. Furthermore,* $\deg p \leq \dim V$.

Proof.

Existence Let $\dim V = n$. We shall induct on n.

If $n = 0$, then I is the zero operator on V. Thus we take p to be the constant polynomial 1.

Suppose $n > 0$, and the desired result holds for all operators on all vector spaces of smaller dimension. We want to construct a monic polynomial of smallest degree such that when applied to T gives the 0 operator.

Let $u \in V \setminus \{\mathbf{0}\}$. Consider the set

$$\{u, Tu, T^2u, \ldots, T^nu\}.$$

This set has length $n+1$, so it is linearly dependent. By the linear dependence lemma, there exists a smallest positive integer $m \leq n$ such that T^mu is a linear combination of $u, Tu, \ldots, T^{m-1}u$; thus there exist $c_i \in \mathbf{F}$ such that

$$c_0u + c_1Tu + \cdots + c_{m-1}T^{m-1}u + T^mu = \mathbf{0}.$$

Define a monic polynomial $q \in \mathbf{F}[z]$ by

$$q(z) = c_0 + c_1 z + \cdots + c_{m-1} z^{m-1} + z^m.$$

Then $q(T)u = \mathbf{0}$. Thus for non-negative integer k,

$$q(T)\left(T^k u\right) = T^k\left(q(T)u\right) = T^k(\mathbf{0}) = \mathbf{0}.$$

By the linear dependence lemma, $\{u, Tu, \ldots, T^{m-1}u\}$ is linearly independent. Thus the above equation implies that $\dim \ker q(T) \geq m$. Hence by the fundamental theorem of linear maps,

$$\begin{aligned} \dim \operatorname{im} q(T) &= \dim V - \dim \ker q(T) \\ &\leq \dim V - m. \end{aligned}$$

Since $\operatorname{im} q(T)$ is invariant under T, we can apply the induction hypothesis to the restriction $T|_{\operatorname{im} q(T)}$. Thus there exists a monic polynomial $s \in \mathbf{F}[z]$ with $\deg s \leq \dim V - m$ such that

$$s\left(T|_{\operatorname{im} q(T)}\right) = 0.$$

Hence for all $v \in V$,

$$((sq)(T))v = s(T)(q(T)v) = \mathbf{0}$$

because $q(T)v \in \operatorname{im} q(T)$ and $s(T)|_{\operatorname{im} q(T)} = s(T|_{\operatorname{im} q(T)}) = 0$. Thus sq is a monic polynomial such that $\deg sq \leq \dim V$ and $(sq)(T) = 0$, as desired.

Uniqueness Let $p \in \mathbf{F}[z]$ be a monic polynomial of smallest degree such that $p(T) = 0$; let $r \in \mathbf{F}[z]$ be a monic polynomial of same degree and $r(T) = 0$. Then $(p-r)(T) = 0$ and also $\deg(p-r) < \deg p$.

We claim that $p - r = 0$. Suppose, for a contradiction, that $p - r \neq 0$. Then divide $p - r$ by the coefficient of the highest-order term in $p - r$ to get a monic polynomial $s \in \mathbf{F}[z]$, which satisfies $s(T) = 0$ and also $\deg s = \deg(p-r) < \deg p$, a contradiction. □

Definition 6.13 (Minimal polynomial)**.** Suppose V is finite-dimensional, $T \in \mathscr{L}(V)$. The ***minimal polynomial*** of T is the unique monic polynomial $p \in \mathbf{F}[z]$ of smallest degree such that $p(T) = 0$.

The minimal polynomial of an operator provides a convenient method to determine its eigenvalues.

Theorem 6.14. *Suppose V is finite-dimensional, $T \in \mathscr{L}(V)$. Then the zeros of the minimal polynomial of T are eigenvalues of T.*

Proof. Let p be the minimal polynomial of T.

$\boxed{\Longrightarrow}$ Let $\lambda \in \mathbf{F}$ be a zero of p. By the factor theorem,

$$p(z) = (z-\lambda)q(z)$$

where q is a monic polynomial with coefficients in $\mathbf{F}$. Since $p(T) = 0$,

$$\mathbf{0} = (T-\lambda I)(q(T)v) \quad (v \in V).$$

Since $\deg p < \deg p$ and p is the minimal polynomial of T, there exists at least one $v \in V$ such that $q(T)v \neq \mathbf{0}$. Thus the equation above implies that λ is an eigenvalue of T.

$\boxed{\Longleftarrow}$ Let $\lambda \in \mathbf{F}$ be an eigenvalue of T. Then there exists $v \in V \setminus \{\mathbf{0}\}$ such that $Tv = \lambda v$. Repeated applications of T to both sides gives $T^k v = \lambda^k v$ for every non-negative integer k. Thus

$$p(T)v = p(\lambda)v.$$

Since p is the minimal polynomial of T, $p(T)v = \mathbf{0}$. Thus the equation above implies that $p(\lambda) = 0$. Hence λ is a zero of p. □

If V is a complex vector space, by the fundamental theorem of algebra (5.11), the minimal polynomial of T has the factorisation

$$(z-\lambda_1)\cdots(z-\lambda_m), \tag{6.1}$$

where $\lambda_1, \ldots, \lambda_m$ are eigenvalues of T (possibly with repetitions).

The next result completely characterises the polynomials that when applied to an operator give the 0 operator.

Proposition 6.15. *Suppose V is finite-dimensional, $T \in \mathscr{L}(V)$, $q \in \mathbf{F}[z]$. Then*

$$q(T) = 0 \iff p \mid q$$

where p is the minimal polynomial of T.

Proof.

$\boxed{\Longrightarrow}$ Suppose $q(T) = 0$. By the division algorithm, there exist polynomials $s, r \in \mathbf{F}[z]$ such that

$$q = ps + r \tag{I}$$

and $\deg r < \deg p$. Then

$$0 = q(T) = p(T)s(T) + r(T) = r(T).$$

The equation above implies that $r = 0$ (otherwise, dividing r by its highest-degree coefficient

would produce a monic polynomial that when applied to T gives 0; this polynomial would have a smaller degree than the minimal polynomial, a contradiction).

Thus (I) becomes $q = ps$, so q is a polynomial multiple of p.

$\boxed{\Longleftarrow}$ Suppose q is a polynomial multiple of p. Then $q = ps$ for some polynomial $s \in \mathbf{F}[z]$, so

$$q(T) = p(T)s(T) = 0s(T) = 0$$

as desired. □

The following corollary concerns the minimal polynomial of a restriction operator.

Corollary 6.16. *Suppose V is finite-dimensional, $T \in \mathscr{L}(V)$, $U \leq V$ is invariant under T. Then the minimal polynomial of T is a polynomial multiple of the minimal polynomial of $T|_U$.*

Proof. Let p be the minimal polynomial of T. Then $p(T)v = \mathbf{0}$ for all $v \in V$. In particular,

$$p(T)u = \mathbf{0} \quad (u \in U).$$

Thus $p(T|_U) = 0$. By 6.15 (applied to $T|_U$ in place of T), p is a polynomial multiple of the minimal polynomial of $T|_U$. □

The next result shows that the constant term of the minimal polynomial determines whether the operator is invertible.

Corollary 6.17. *Suppose V is finite-dimensional, $T \in \mathscr{L}(V)$. Then T is not invertible if and only if the constant term of the minimal polynomial of T is 0.*

Proof. Suppose $T \in \mathscr{L}(V)$, let p be the minimal polynomial of T. Then

$$\begin{aligned} T \text{ is not invertible} &\iff 0 \text{ is an eigenvalue of } T && \text{[by 6.4]} \\ &\iff 0 \text{ is a zero of } p && \text{[by 6.14]} \\ &\iff \text{constant term of } p \text{ is } 0. \end{aligned}$$

□

Eigenvalues on Odd-Dimensional Real Vector Spaces

The next result will be the key tool that we use to show that every operator on an odd-dimensional real vector space has an eigenvalue.

Lemma 6.18. *Suppose V is a finite-dimensional, real vector space. Suppose $T \in \mathscr{L}(V)$ and $b, c \in \mathbb{R}$ with $b^2 < 4c$. Then $\dim \ker(T^2 + bT + cI)$ is even.*

Proof. By 6.10, $\ker(T^2 + bT + cI)$ is invariant under T. By replacing V with $\ker(T^2 + bT + cI)$ and replacing T with $T|_{\ker(T^2+bT+cI)}$, we can assume that $T^2 + bT + cI = 0$; we now need to prove that $\dim V$ is even.

Claim. T has no eigenvectors.

Suppose $\lambda \in \mathbb{R}$ and $v \in V$ are such that $Tv = \lambda v$. Then

$$\mathbf{0} = (T^2 + bT + cI)v = (\lambda^2 + b\lambda + c)v = \underbrace{\left(\left(\lambda + \frac{b}{2}\right)^2 + c - \frac{b^2}{4}\right)}_{>0} v$$

implies $v = \mathbf{0}$. Hence T has no eigenvectors.

Let $U \leq V$ be invariant under T, and has the largest dimension among all subspaces of V that are invariant under T and have even dimension.

Claim. $U = V$.

Suppose, for a contradiction, that $U \neq V$. Then there exists $w \in V$ such that $w \notin U$.

Let $W = \operatorname{span}(w, Tw)$. Then W is invariant under T, since $T(Tw) = -bTw - cw$. Furthermore, $\dim W = 2$, since otherwise w would be an eigenvector of T. Now

$$\dim(U + W) = \dim U + \dim W - \dim(U \cap W) = \dim U + 2 \tag{I}$$

where $U \cap W = \{\mathbf{0}\}$ because otherwise $U \cap W$ would be a one-dimensional subspace of V that is invariant under T (impossible because T has no eigenvectors).

Since $U + W$ is invariant under T, (I) shows that there exists a subspace of V invariant under T of even dimension larger than $\dim U$. Thus the assumption that $U \neq V$ was incorrect. Hence V has even dimension. □

The next result states that on odd-dimensional vector spaces, every operator has an eigenvalue. We already know this result for finite-dimensional complex vectors spaces (without the odd hypothesis). Thus in the proof below, we will assume that $\mathbf{F} = \mathbb{R}$.

Theorem 6.19. *Every operator on an odd-dimensional vector space has an eigenvalue.*

Proof. Suppose V is a finite-dimensional real vector space, $\dim V = n$ is odd. Let $T \in \mathscr{L}(V)$. We will induct on n in steps of size two to show that T has an eigenvalue.

If $n = 1$, then every non-zero vector in V is an eigenvector of T. Thus the desired result holds.

Suppose $n \geq 3$, and the desired result holds for all operators on all odd-dimensional vector spaces of dimension less than n.

Let p be the minimal polynomial of T. If p is a polynomial multiple of $x - \lambda$ for some $\lambda \in \mathbb{R}$, by 6.14, λ is an eigenvalue of T and we are done. Thus we can assume that there exist $b, c \in \mathbb{R}$ such that $b^2 < 4c$ and p is a polynomial multiple of $x^2 + bx + c$ (see 5.14). There exists a monic polynomial $q \in \mathbb{R}[x]$ such that

$$p(x) = q(x)(x^2 + bx + c) \quad (x \in \mathbb{R}).$$

Then

$$0 = p(T) = (q(T))(T^2 + bT + cI),$$

which means that $q(T)$ equals 0 on $\operatorname{im}(T^2 + bT + cI)$. Since $\deg q < \deg p$ and p is the minimal polynomial of T, this implies that $\operatorname{im}(T^2 + bT + cI) \neq V$.

By the fundamental theorem of linear maps,

$$\underbrace{\dim V}_{\text{odd}} = \underbrace{\dim \ker(T^2 + bT + cI)}_{\text{even}} + \dim \operatorname{im}(T^2 + bT + cI)$$

implies that $\dim \operatorname{im}(T^2 + bT + cI)$ is odd.

Hence $\operatorname{im}(T^2 + bT + cI)$ is a subspace of V that is invariant under T (by 6.10) and that has odd dimension less than $\dim V$. By induction hypothesis, $T|_{\operatorname{im}(T^2+bT+cI)}$ has an eigenvalue, which means that T has an eigenvalue. □

6.3 Upper-Triangular Matrices

Suppose $T \in \mathscr{L}(V)$. Recall that the matrix of T with respect to a basis $\{v_1, \ldots, v_n\}$ of V is the $n \times n$ matrix whose entries A_{ij} are defined by

$$Tv_j = \sum_{i=1}^{n} A_{ij} v_i \quad (j = 1, \ldots, n).$$

Notation. If the basis is not clear from context, we denote the matrix of T as $\mathscr{M}(T; \{v_1, \ldots, v_n\})$.

Remark. The matrices of operators are square matrices.

The *diagonal* of a square matrix consists of the entries on the line from the upper left corner to the bottom right corner.

Definition 6.20 (Upper-triangular matrix)**.** A square matrix is called ***upper-triangular*** if all the entries below the diagonal are 0.

We represent an upper-triangular matrix in the form

$$\begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

where the 0 indicates that all entries below the diagonal equal 0, and $*$ denotes entries that we do not know or that are irrelevant to the questions being discussed.

The next result provides a useful connection between upper-triangular matrices and invariant subspaces.

Lemma 6.21 (Conditions for upper-triangular matrix)**.** *Suppose $T \in \mathscr{L}(V)$, and $\{v_1, \ldots, v_n\}$ is a basis of V. Then the following are equivalent:*

(i) $\mathscr{M}(T; \{v_1, \ldots, v_n\})$ *is upper-triangular.*

(ii) $\operatorname{span}(v_1, \ldots, v_k)$ *is invariant under T for each $k = 1, \ldots, n$.*

(iii) $Tv_k \in \operatorname{span}(v_1, \ldots, v_k)$ *for each $k = 1, \ldots, n$.*

Proof.

(i) $\Longrightarrow$ (ii) Let $k \in \{1, \ldots, n\}$. Since the matrix of T with respect to $\{v_1, \ldots, v_n\}$ is upper-triangular, if $j \in \{1, \ldots, n\}$, then

$$Tv_j = \sum_{i=1}^{n} A_{ij} v_i = \sum_{i=1}^{j} A_{ij} v_i + 0v_{j+1} + \cdots + 0v_n \in \operatorname{span}(v_1, \ldots, v_j).$$

If $j \le k$, then $\operatorname{span}(v_1,\dots,v_j) \subset \operatorname{span}(v_1,\dots,v_k)$, so

$$Tv_j \in \operatorname{span}(v_1,\dots,v_k)$$

for each $j \in \{1,\dots,k\}$. Hence $\operatorname{span}(v_1,\dots,v_k)$ is invariant under T.

$\boxed{\text{(ii)} \Longrightarrow \text{(iii)}}$ Suppose $\operatorname{span}(v_1,\dots,)$ is invariant under T for each $k = 1,\dots,n$.

In particular, $Tv_k \in \operatorname{span}(v_1,\dots,v_k)$ for each $k = 1,\dots,n$.

$\boxed{\text{(iii)} \Longrightarrow \text{(i)}}$ Suppose $Tv_k \in \operatorname{span}(v_1,\dots,v_k)$ for each $k = 1,\dots,n$.

Then when writing each Tv_k as a linear combination of basis vectors $v_1,\dots,v_n$, we need to use only $v_1,\dots,v_k$. Hence all entries under the diagonal of $\mathscr{M}(T)$ are 0, so $\mathscr{M}(T)$ is an upper-triangular matrix. □

The next result tells us that if $\mathscr{M}(T)$ is upper-triangular with respect to some basis of V, then T satisfies a simple equation depending on the diagonal entries.

Proposition 6.22. *Suppose $T \in \mathscr{L}(V)$ has an upper-triangular matrix with respect to some basis of V. If the diagonal entries are $\lambda_1,\dots,\lambda_n$, then*

$$(T-\lambda_1 I)\cdots(T-\lambda_n I) = 0. \tag{6.2}$$

Proof. Suppose T has an upper-triangular matrix with respect to the basis $\{v_1,\dots,v_n\}$ of V. Let the diagonal entries be $\lambda_1,\dots,\lambda_n$; that is,

$$\mathscr{M}(T) = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

- Considering the first column of $\mathscr{M}(T)$, we have

$$\begin{aligned} & Tv_1 = \lambda_1 v_1 \\ \Longrightarrow\ & (T-\lambda_1 I)v_1 = \mathbf{0} \\ \Longrightarrow\ & (T-\lambda_1 I)\cdots(T-\lambda_m I)v_1 = \mathbf{0} \quad (m = 1,\dots,n). \end{aligned}$$

- Considering the second column of $\mathscr{M}(T)$, we have

$$\begin{aligned} & (T-\lambda_2 I)v_2 \in \operatorname{span}(v_1) \\ \Longrightarrow\ & (T-\lambda_1 I)(T-\lambda_2 I)v_2 = \mathbf{0} \\ \Longrightarrow\ & (T-\lambda_1 I)\cdots(T-\lambda_m I)v_2 = \mathbf{0} \quad (m = 2,\dots,n). \end{aligned}$$

- Considering the third column of $\mathscr{M}(T)$, we have

$$\begin{aligned} &(T-\lambda_3 I)v_3 \in \operatorname{span}(v_1, v_2) \\ \Longrightarrow\ &(T-\lambda_1 I)(T-\lambda_2 I)(T-\lambda_3 I)v_3 = \mathbf{0} \\ \Longrightarrow\ &(T-\lambda_1 I)\cdots(T-\lambda_m I)v_3 = \mathbf{0} \quad (m = 3,\ldots,n). \end{aligned}$$

Continuing this pattern, we see that

$$(T-\lambda_1 I)\cdots(T-\lambda_n I)v_k = \mathbf{0} \quad (k = 1,\ldots,n).$$

Hence $(T-\lambda_1 I)\cdots(T-\lambda_n I)$ is the 0 operator, because it is $\mathbf{0}$ on each vector in a basis of V. □

The next result tells us that the eigenvalues of an operator can be determined from the upper-triangular matrix.

Proposition 6.23. *Suppose $T \in \mathscr{L}(V)$ has an upper-triangular matrix with respect to some basis of V. Then the eigenvalues of T are precisely the entries on the diagonal of that upper-triangular matrix.*

Proof. Let $\{v_1,\ldots,v_n\}$ be a basis of V with respect to which T has an upper-triangular matrix:

$$\mathscr{M}(T) = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

Since $Tv_1 = \lambda_1 v_1$, we see that λ_1 is an eigenvalue of T.

Let $k \in \{2,\ldots,n\}$, then $(T-\lambda_k I)v_k \in \operatorname{span}(v_1,\ldots,v_{k-1})$, so $T-\lambda_k I$ maps $\operatorname{span}(v_1,\ldots,v_k)$ into $\operatorname{span}(v_1,\ldots,v_{k-1})$. Since

$$\dim \operatorname{span}(v_1,\ldots,v_k) = k, \quad \dim \operatorname{span}(v_1,\ldots,v_{k-1}) = k-1,$$

this implies that $T-\lambda_k I$ restricted to $\operatorname{span}(v_1,\ldots,v_k)$ is not injective (by 4.12). Thus there exists $v \in \operatorname{span}(v_1,\ldots,v_k)$ such that $v \neq \mathbf{0}$ and $(T-\lambda_k I)v = \mathbf{0}$. Thus λ_k is an eigenvalue of T. Hence every entry on the diagonal of $\mathscr{M}(T)$ is an eigenvalue of T.

We now prove T has no other eigenvalues. Let q be the polynomial defined by

$$q(z) = (z-\lambda_1)\cdots(z-\lambda_n).$$

By 6.22, $q(T) = 0$. By 6.15, q is a polynomial multiple of the minimal polynomial of T. Thus every zero of the minimal polynomial of T is a zero of q.

By 6.14, the zeros of the minimal polynomial of T are the eigenvalues of T. This implies that every eigenvalue of T is a zero of q.

Hence the eigenvalues of T are all contained in the set $\{\lambda_1, \dots, \lambda_n\}$. □

Example. Define $T \in \mathscr{L}(\mathbf{F}^3)$ by $T(x,y,z) = (2x+y, 5y+3z, 8z)$. The matrix of T with respect to the standard basis is

$$\mathscr{M}(T) = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 5 & 3 \\ 0 & 0 & 8 \end{pmatrix}.$$

Thus the eigenvalues of T are 2, 5, and 8.

The following result gives a *necessary and sufficient condition* to have an upper-triangular matrix.

Proposition 6.24. *Suppose V is finite-dimensional, $T \in \mathscr{L}(V)$. Then T has an upper-triangular matrix with respect to some basis of V if and only if the minimal polynomial equals*

$$(z-\lambda_1)\cdots(z-\lambda_m)$$

for some $\lambda_1, \dots, \lambda_m \in \mathbf{F}$.

Proof.

$\boxed{\Longrightarrow}$ Suppose T has an upper-triangular matrix with respect to some basis of V. Let $\alpha_1, \dots, \alpha_n$ denote the diagonal entries of that matrix. Define a polynomial $q \in \mathbf{F}[z]$ by

$$q(z) = (z-\alpha_1)\cdots(z-\alpha_n).$$

By 6.22, $q(T) = 0$. By 6.15, q is a polynomial multiple of the minimal polynomial. Thus the minimal polynomial of T equals $(z-\lambda_1)\cdots(z-\lambda_m)$ for some $\lambda_1, \dots, \lambda_m \in \mathbf{F}$ with $\{\lambda_1, \dots, \lambda_m\} \subset \{\alpha_1, \dots, \alpha_n\}$.

$\boxed{\Longleftarrow}$ Suppose the minimal polynomial of T equals $(z-\lambda_1)\cdots(z-\lambda_m)$ for some $\lambda_1, \dots, \lambda_m \in \mathbf{F}$. We induct on m.

For the base case $m = 1$, $z-\lambda_1$ is the minimal polynomial of T, which implies that $T = \lambda_1 I$, so the matrix of T (with respect to any basis of V) is upper-triangular.

Now suppose $m > 1$ and the desired result holds for all smaller positive integers. Let

$$U = \operatorname{im}(T - \lambda_m I).$$

Then U is invariant under T (by 6.10), so $T|_U$ is an operator on U.

If $u \in U$, then $u = (T-\lambda_m I)v$ for some $v \in V$ and

$$(T-\lambda_1 I)\cdots(T-\lambda_{m-1} I)u = (T-\lambda_1 I)\cdots(T-\lambda_m I)v = \mathbf{0}.$$

Hence $(z-\lambda_1)\cdots(z-\lambda_{m-1})$ is a polynomial multiple of the minimal polynomial of $T|_U$, by 6.15. Thus the minimal polynomial of $T|_U$ is the product of at most $m-1$ terms of the form $z-\lambda_k$.

By our induction hypothesis, there is a basis $\{u_1,\ldots,u_M\}$ of U, with respect to which $T|_U$ has an upper-triangular matrix. Thus for each $k\in\{1,\ldots,M\}$, we have (using 6.21)

$$Tu_k = (T|_U)(u_k) \in \operatorname{span}(u_1,\ldots,u_k). \tag{I}$$

Extend $\{u_1,\ldots,u_M\}$ to a basis $\{u_1,\ldots,u_M,v_1,\ldots,v_N\}$ of V. If $k\in\{1,\ldots,N\}$, then

$$Tv_k = (T-\lambda_m I)v_k + \lambda_m v_k.$$

The definition of U shows that $(T-\lambda_m I)v_k \in U = \operatorname{span}(u_1,\ldots,u_M)$. Thus the equation above shows that

$$Tv_k \in \operatorname{span}(u_1,\ldots,u_M,v_1,\ldots,v_k). \tag{II}$$

From (I) and (II), we conclude (using 6.21) that T has an upper-triangular matrix with respect to the basis $\{u_1,\ldots,u_M,v_1,\ldots,v_N\}$ of V, as desired. □

We conclude with an important result: every operator on a finite-dimensional complex vector space has an upper-triangular matrix.

Theorem 6.25. *Suppose V is finite-dimensional complex vector space, $T\in\mathscr{L}(V)$. Then T has an upper-triangular matrix with respect to some basis of V.*

Proof. By the fundamental theorem of algebra II (5.11), the minimal polynomial of T has a factorisation

$$p(z) = (z-\lambda_1)\cdots(z-\lambda_m)$$

for some $\lambda_1,\ldots,\lambda_m\in\mathbb{C}$. By 6.24, T has an upper-triangular matrix with respect to some basis of V. □

6.4 Diagonalisable Operators

Diagonal Matrices

Definition 6.26. A ***diagonal matrix*** is a square matrix that is 0 everywhere except possibly on the diagonal.

That is, a diagonal matrix has the form

$$\begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

for some $\lambda_1, \ldots, \lambda_n \in \mathbf{F}$.

Remark. If an operator has a diagonal matrix with respect to some basis, then the entries on the diagonal are precisely the eigenvalues of the operator, by 6.23.

Definition 6.27. Suppose $T \in \mathscr{L}(V)$, $\lambda \in \mathbf{F}$. The ***eigenspace*** of T corresponding to λ is

$$E(\lambda, T) := \ker(T - \lambda I) = \{v \in V \mid Tv = \lambda v\}.$$

Hence $E(\lambda, T)$ is the set of all eigenvectors of T corresponding to λ, along with the $\mathbf{0}$ vector. The definitions imply that λ is an eigenvalue of T if and only if $E(\lambda, T) \neq \{\mathbf{0}\}$.

By 6.10, $E(\lambda, T)$ is a subspace of V.

Remark. If λ is an eigenvalue of $T \in \mathscr{L}(V)$, then $T|_{E(\lambda,T)}$ is the operator of multiplication by λ.

The next result states that the sum of eigenspaces is a direct sum.

Proposition 6.28. *Suppose $T \in \mathscr{L}(V)$, and $\lambda_1, \ldots, \lambda_m$ are distinct eigenvalues of T. Then*

$$E(\lambda_1, T) \oplus \cdots \oplus E(\lambda_m, T).$$

Proof. To show that $E(\lambda_1, T) + \cdots + E(\lambda_m, T)$ is a direct sum, suppose

$$v_1 + \cdots + v_m = \mathbf{0},$$

where each $v_i \in E(\lambda_i, T)$. By 6.5, eigenvectors corresponding to distinct eigenvalues are linearly independent, so each $v_i = \mathbf{0}$.

Hence by 3.12, $E(\lambda_1, T) + \cdots + E(\lambda_m, T)$ is a direct sum. □

If V is finite-dimensional, then

$$\begin{aligned} \dim E(\lambda_1,T)+\cdots+\dim E(\lambda_m,T) &= \dim\left(E(\lambda_1,T)\oplus\cdots\oplus E(\lambda_m,T)\right) && \text{[by 4.50]} \\ &\leq \dim V && \text{[by 3.31]} \end{aligned}$$

Conditions for Diagonalisability

Definition 6.29 (Diagonalisable)**.** Suppose $T\in\mathscr{L}(V)$. We say T is ***diagonalisable*** if T has a diagonal matrix with respect to some basis of V.

Diagonalisation may require a different basis.

Example. Define $T\in\mathscr{L}(\mathbb{R}^2)$ by

$$T(x,y)=(41x+7y,-20x+74y).$$

The matrix of T is respect to the standard basis of $\mathbb{R}^2$ is

$$\begin{pmatrix} 41 & 7 \\ -20 & 74 \end{pmatrix}$$

which is not a diagonal matrix. However T is diagonalisable, because the matrix of T with respect to the basis $(1,4),(7,5)$ is

$$\begin{pmatrix} 69 & 0 \\ 0 & 46 \end{pmatrix}.$$

The following characterisations of diagonalisable operators will be useful.

Lemma 6.30 (Conditions equivalent to diagonalisability)**.** *Suppose V is finite-dimensional, $T\in\mathscr{L}(V)$, and $\lambda_1,\ldots,\lambda_m$ are distinct eigenvalues of T. Then the following are equivalent:*

(i) *T is diagonalisable.*

(ii) *V has a basis consisting of eigenvectors of T.*

(iii) $V=E(\lambda_1,T)\oplus\cdots\oplus E(\lambda_m,T)$. *(direct sum decomposition)*

(iv) $\dim V=\dim E(\lambda_1,T)+\cdots+\dim E(\lambda_m,T)$.

Proof.

(i) $\iff$ (ii) By definition, $T \in \mathscr{L}(V)$ is diagonalisable if and only if T has a diagonal matrix

$$\begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

with respect to a basis $\{v_1, \ldots, v_n\}$ of V, if and only if $Tv_i = \lambda_i v_i$ for each $i = 1, \ldots, n$. Hence $v_1, \ldots, v_n$ are eigenvectors of T.

(ii) $\implies$ (iii) Suppose V has a basis consisting of eigenvectors of T. Then every vector in V can be written as a linear combination of eigenvectors of T, which implies that

$$V = E(\lambda_1, T) + \cdots + E(\lambda_m, T).$$

By 6.28, this is a direct sum.

(iii) $\implies$ (iv) This follows from 4.50.

(iv) $\implies$ (ii) Suppose

$$\dim V = \dim E(\lambda_1, T) + \cdots + \dim E(\lambda_m, T).$$

Choose a basis of each $E(\lambda_i, T)$; put all these bases together to form a set $\{v_1, \ldots, v_n\}$ of eigenvectors of T, where $n = \dim V$.

Claim. $\{v_1, \ldots, v_n\}$ is a basis of T.

By 3.32, it suffices to show that $\{v_1, \ldots, v_n\}$ is linearly independent. Suppose $a_1, \ldots, a_n \in \mathbf{F}$ are such that

$$a_1 v_1 + \cdots + a_n v_n = \mathbf{0}.$$

For each $i = 1, \ldots, m$, let u_i denote the sum of all the terms $a_j v_j$ such that $v_j \in E(\lambda_i, T)$. Thus each u_i is in $E(\lambda_i, T)$, and

$$u_1 + \cdots + u_m = \mathbf{0}.$$

By 6.5, since eigenvectors corresponding to distinct eigenvalues are linearly independent, this implies that each u_i equals $\mathbf{0}$. Because each u_i is a sum of terms $a_j v_j$, where the v_j's were chosen to be a basis of $E(\lambda_i, T)$, this implies that all a_j's equal 0. Thus $\{v_1, \ldots, v_n\}$ is linearly independent and hence is a basis of V. □

The next result shows that if an operator has as many distinct eigenvalues as the dimension of its domain, then the operator is diagonalisable.

Corollary 6.31. *Suppose V is finite-dimensional, $T \in \mathscr{L}(V)$ has $\dim V$ distinct eigenvalues. Then T is diagonalisable.*

Proof. Suppose T has distinct eigenvalues $\lambda_1, \ldots, \lambda_{\dim V}$, with corresponding eigenvectors

$v_1, \ldots, v_{\dim V}$. By 6.5, eigenvectors corresponding to distinct eigenvalues are linearly independent, so $v_1, \ldots, v_{\dim V}$ are linearly independent, and thus forms a basis of V.

With respect to this basis consisting of eigenvectors, T has a diagonal matrix. □

In later chapters, we will find additional conditions that imply that certain operators are diagonalisable. For example, see the real spectral theorem (8.19) and the complex spectral theorem (8.20).

The next result provides a necessary and sufficient condition for diagonalisability.

Theorem 6.32. *Suppose V is finite-dimensional, $T \in \mathscr{L}(V)$. Then T is diagonalisable if and only if the minimal polynomial of T equals*

$$(z-\lambda_1)\cdots(z-\lambda_m)$$

for distinct $\lambda_1, \ldots, \lambda_m \in \mathbf{F}$.

Proof.

$\boxed{\Longrightarrow}$ Suppose T is diagonalisable. Thus there is a basis $\{v_1, \ldots, v_n\}$ of V consisting of eigenvectors of T. Let $\lambda_1, \ldots, \lambda_m$ be distinct eigenvalues of T. Then for each v_j, there exists λ_k with $(T-\lambda_k I)v_j = \mathbf{0}$. Thus

$$(T-\lambda_1 I)\cdots(T-\lambda_m I)v_j = \mathbf{0},$$

which implies that the minimal polynomial of T equals $(z-\lambda_1)\cdots(z-\lambda_m)$.

$\boxed{\Longleftarrow}$ Suppose the minimal polynomial of T equals $(z-\lambda_1)\cdots(z-\lambda_m)$ for distinct $\lambda_1, \ldots, \lambda_m \in \mathbf{F}$. Thus

$$(T-\lambda_1 I)\cdots(T-\lambda_m I) = 0. \tag{I}$$

We will prove that T is diagonalisable by inducting on m. For the base case $m = 1$, $T - \lambda_1 I = 0$, which means that T is a scalar multiple of the identity operator, so T is diagonalisable.

Now suppose that $m > 1$ and the desired result holds for all smaller values of m. The subspace $\operatorname{im}(T-\lambda_m I)$ is invariant under T (by 6.10). Thus $T|_{\operatorname{im}(T-\lambda_m I)} \in \mathscr{L}(\operatorname{im}(T-\lambda_m I))$.

If $u \in \operatorname{im}(T-\lambda_m I)$, then $u = (T-\lambda_m I)v$ for some $v \in V$, and (I) implies

$$(T-\lambda_1 I)\cdots(T-\lambda_{m-1} I)u = (T-\lambda_1)\cdots(T-\lambda_m I)v = \mathbf{0}. \tag{II}$$

Hence $(z-\lambda_1)\cdots(z-\lambda_{m-1})$ is a polynomial multiple of the minimal polynomial of $T|_{\operatorname{im}(T-\lambda_m I)}$, by 6.15. Thus by induction hypothesis, there is a basis of $\operatorname{im}(T-\lambda_m I)$ consisting of eigenvectors of T.

Let $u \in \operatorname{im}(T-\lambda_m I) \cap \ker(T-\lambda_m I)$. Then $Tu = \lambda_m u$. Now (II) implies that

$$\begin{aligned} \mathbf{0} &= (T-\lambda_1 I)\cdots(T-\lambda_{m-1}I)u \\ &= (\lambda_m-\lambda_1)\cdots(\lambda_m-\lambda_{m-1})u. \end{aligned}$$

Since $\lambda_1, \ldots, \lambda_m$ are distinct, the equation above implies that $u = \mathbf{0}$. Hence $\operatorname{im}(T-\lambda_m I) \cap \ker(T-\lambda_m I) = \{\mathbf{0}\}$.

By 3.13, this implies that $\operatorname{im}(T-\lambda_m I) + \ker(T-\lambda_m I)$ is a direct sum, whose dimension is $\dim V$ (by 4.50 and the fundamental theorem of linear maps). Hence $\operatorname{im}(T-\lambda_m I) \oplus \ker(T-\lambda_m I) = V$. Every nonzero vector in $\ker(T-\lambda_m I)$ is an eigenvector of T with eigenvalue λ_m. Earlier in this proof we saw that there is a basis of $\operatorname{im}(T-\lambda_m I)$ consisting of eigenvectors of T. Adjoining to that basis a basis of $\ker(T-\lambda_m I)$ gives a basis of V consisting of eigenvectors of T. The matrix of T with respect to this basis is a diagonal matrix, as desired. □

The next result states that the restriction of a diagonalisable operator to an invariant subspace is still diagonalisable.

Corollary 6.33. *Suppose $T \in \mathscr{L}(V)$ is diagonalisable, $U \leq V$ is invariant under T. Then $T|_U$ is a diagonalisable operator on U.*

Proof. Since T is diagonalisable, by 6.32, the minimal polynomial of T equals $(z-\lambda_1)\cdots(z-\lambda_m)$ for some distinct $\lambda_1, \ldots, \lambda_m \in \mathbf{F}$.

By 6.16, the minimal polynomial of T is a polynomial multiple of the minimal polynomial of $T|_U$. Hence the minimal polynomial of $T|_U$ has the form required by 6.32, which shows that $T|_U$ is diagonalisable. □

6.5 Commuting Operators

Definition 6.34 (Commute). Two operators S and T on the same vector space ***commute*** if $ST = TS$.
Two square matrices A and B of the same size *commute* if $AB = BA$.

The next result shows that two operators commute if and only if their matrices (with respect to the same basis) commute.

Lemma 6.35. *Suppose $S, T \in \mathscr{L}(V)$ and $\{v_1, \ldots, v_n\}$ is a basis of V. Then S and T commute if and only if $\mathscr{M}(S; \{v_1, \ldots, v_n\})$ and $\mathscr{M}(T; \{v_1, \ldots, v_n\})$ commute.*

Proof. We have

$$\begin{aligned} S \text{ and } T \text{ commute} &\iff ST = TS \\ &\iff \mathscr{M}(ST) = \mathscr{M}(TS) \\ &\iff \mathscr{M}(S)\mathscr{M}(T) = \mathscr{M}(T)\mathscr{M}(S) \\ &\iff \mathscr{M}(S) \text{ and } \mathscr{M}(T) \text{ commute} \end{aligned}$$

as desired. □

The next result shows that if two operators commute, then every eigenspace for one operator is invariant under the other operator.

Lemma 6.36. *Suppose $S, T \in \mathscr{L}(V)$ commute, $\lambda \in \mathbf{F}$. Then $E(\lambda, S)$ is invariant under T.*

Proof. Let $v \in E(\lambda, S)$. Then

$$S(Tv) = (ST)v = (TS)v = T(Sv) = T(\lambda v) = \lambda Tv \implies Tv \in E(\lambda, S).$$

Hence $E(\lambda, S)$ is invariant under T. □

If we want to do computations with two diagonalisable operators, we would want them to be diagonalisable by the same basis. The next result states that this is possible when the two operators commute.

Proposition 6.37. *Two diagonalisable operators on the same vector space have diagonal matrices with respect to the same basis if and only if the two operators commute.*

Proof.

$\boxed{\Longrightarrow}$ Suppose $S, T \in \mathscr{L}(V)$ have diagonal matrices with respect to the same basis.

Since any two diagonal matrices of the same size commute, by 6.35, S and T commute.

$\boxed{\Longleftarrow}$ Suppose $S, T \in \mathscr{L}(V)$ are diagonalisable operators that commute, so $ST = TS$. Let $\lambda_1, \ldots, \lambda_m$ denote the distinct eigenvalues of S.

Since S is diagonalisable, by 6.30,

$$V = E(\lambda_1, S) \oplus \cdots \oplus E(\lambda_m, S). \tag{I}$$

For each $i = 1, \ldots, m$, the subspace $E(\lambda_i, S)$ is invariant under T (by 6.36). Since T is diagonalisable, by 6.33, the restriction $T|_{E(\lambda_i, S)}$ is diagonalisable for each i.

Hence for each $i = 1, \ldots, m$, there is a basis of $E(\lambda_i, S)$ consisting of eigenvectors of T. Putting these bases together gives a basis of V (because of (I)), with each vector in this basis being an eigenvector of both S and T. Thus S and T both have diagonal matrices with respect to this basis, as desired. □

Suppose V is a finite-dimensional nonzero complex vector space. Then every operator on V has an eigenvector (by 6.11). The next result shows that if two operators on V commute, then there is a vector in V that is an eigenvector for both operators (but the two commuting operators might not have a common eigenvalue).

Lemma 6.38. *Every pairs of commuting operators on a finite-dimensional non-zero complex vector space has a common eigenvector.*

Proof. Suppose V is a finite-dimensional nonzero complex vector space. Suppose $S, T \in \mathscr{L}(V)$ commute.

Let λ be an eigenvalue of S (6.11 tells us that S does indeed have an eigenvalue). Thus $E(\lambda, S) \neq \{\mathbf{0}\}$. Also, $E(\lambda, S)$ is invariant under T (by 6.36).

Thus $T|_{E(\lambda, S)}$ has an eigenvector (again using 6.11), which is an eigenvector for both S and T, completing the proof. □

Remark. The hypothesis $\mathbb{C}$ is needed, since all vector spaces over $\mathbb{C}$ have eigenvalues, by 6.11.

Recall that 6.25 states that for every operator, there exists a basis that gives an upper-triangular matrix. We now extend this result to two commuting operators.

Proposition 6.39. *Suppose V is a finite-dimensional complex vector space, $S, T \in \mathscr{L}(V)$ commute. Then there exists a basis of V, with respect to which both S and T have upper-triangular matrices.*

Proof. Let $n = \dim V$. Induct on n.

The desired result holds if $n = 1$, since all 1×1 matrices are upper-triangular.

Now suppose $n > 1$ and the desired result holds for all complex vector spaces whose dimension is $n-1$.

Since S and T commute, by 6.38, let v_1 be any common eigenvalue of S and T. Hence $Sv_1 \in \operatorname{span}(v_1)$ and $Tv_1 \in \operatorname{span}(v_1)$. By 3.28, there exists a subspace W of V such that

$$V = \operatorname{span}(v_1) \oplus W.$$

Define a linear map $P\colon V \to W$ by

$$P(av_1 + w) = w \quad (a \in \mathbb{C}, w \in W).$$

Define $\tilde{S}, \tilde{T} \in \mathscr{L}(W)$ by

$$\tilde{S}w = P(Sw), \quad \tilde{T}w = P(Tw)$$

for each $w \in W$. To apply the induction hypothesis to $\tilde{S}$ and $\tilde{T}$, we must first show that they commute. Let $w \in W$, then there exists $a \in \mathbb{C}$ such that

$$(\tilde{S}\tilde{T})w = \tilde{S}(P(Tw)) = \tilde{S}(Tw - av_1) = P(S(Tw - av_1)) = P((ST)w),$$

where the last equality holds because v_1 is an eigenvector of S and $Pv_1 = 0$. Similarly,

$$(\tilde{T}\tilde{S})w = P((TS)w).$$

Since S and T commute, the last two displayed equations show that $(\tilde{S}T)w = (\tilde{T}S)w$. Hence $\tilde{S}$ and $\tilde{T}$ commute.

Thus we can use our induction hypothesis to state that there exists a basis $\{v_2, \dots, v_n\}$ of W such that $\tilde{S}$ and $\tilde{T}$ both have upper-triangular matrices with respect to this basis. The list $\{v_1, \dots, v_n\}$ is a basis of V.

If $k \in \{2, \dots, n\}$, then there exist $a_k, b_k \in \mathbb{C}$ such that

$$Sv_k = a_k v_1 + \tilde{S}v_k$$
$$Tv_k = b_k v_1 + \tilde{T}v_k$$

Since $\tilde{S}$ and $\tilde{T}$ have upper-triangular matrices with respect to $\{v_2, \dots, v_n\}$, we know that $\tilde{S}v_k \in \operatorname{span}(v_2, \dots, v_k)$ and $\tilde{T}v_k \in \operatorname{span}(v_2, \dots, v_k)$. Hence the equations above imply that

$$Sv_k \in \operatorname{span}(v_1, \dots, v_k), \quad Tv_k \in \operatorname{span}(v_1, \dots, v_k).$$

Hence S and T have upper-triangular matrices with respect to $\{v_1, \dots, v_n\}$, as desired. □

In general, it is not possible to determine the eigenvalues of the sum or product of two operators from the eigenvalues of the two operators. However, the next result shows that something nice happens when the two operators commute.

Proposition 6.40 (Eigenvalues of sum and product of commuting operators)**.** *Suppose V is a finite-dimensional complex vector space, S and T are commuting operators on V. Then*

(i) every eigenvalue of $S+T$ is an eigenvalue of S plus an eigenvalue of T;

(ii) every eigenvalue of ST is an eigenvalue of S times an eigenvalue of T.

Proof.

(i) By 6.39, there exists a basis of V, with respect to which both S and T have upper-triangular matrices. With respect to that basis,

$$\mathscr{M}(S+T) = \mathscr{M}(S) + \mathscr{M}(T).$$

By definition of matrix addition, each entry on the diagonal of $\mathscr{M}(S+T)$ equals the sum of the corresponding entries on the diagonals of $\mathscr{M}(S)$ and $\mathscr{M}(T)$. Furthermore, $\mathscr{M}(S+T)$ is upper-triangular (as you should verify).

By 6.23,

- every entry on the diagonal of $\mathscr{M}(S)$ is an eigenvalue of S,
- every entry on the diagonal of $\mathscr{M}(T)$ is an eigenvalue of T, and
- every eigenvalue of $S+T$ is on the diagonal of $\mathscr{M}(S+T)$.

Hence every eigenvalue of $S+T$ is an eigenvalue of S plus an eigenvalue of T.

(ii) Similar to above.

□

Exercises

Exercise 6.1 ([Axl24] 5A Q1)**.** Suppose $T \in \mathscr{L}(V)$, $U \leq V$. Prove that

(i) if $U \subset \ker T$, then U is invariant under T;

(ii) if $\operatorname{im} T \subset U$, then U is invariant under T.

Solution.

(i)

(ii) Let $u \in U$. Then $Tu \in \operatorname{im} T \subset U$ so $Tu \in U$.

□

Exercise 6.2 ([Axl24] 5A Q4)**.**

Exercise 6.3 ([Axl24] 5A Q8)**.**

Exercise 6.4 ([Axl24] 5A Q11)**.**

Exercise 6.5 ([Axl24] 5A Q13)**.**

Exercise 6.6 ([Axl24] 5A Q28)**.**

Exercise 6.7 ([Axl24] 5A Q32)**.**

5B 2 7 10 11 13 17 18 22

Exercise 6.8 ([Axl24] 5D Q1)**.** Suppose V is a finite-dimensional complex vector space and $T \in \mathscr{L}(V)$.

(i) Prove that if $T^4 = I$, then T is diagonalisable.

(ii) Prove that if $T^4 = T$, then T is diagonalisable.

(iii) Give an example of an operator $T \in \mathscr{L}(\mathbb{C}^2)$ such that $T^4 = T^2$ and T is not diagonalisable.

Solution.

(i) If $T^4 = I$, then $T^4 - I = 0$. Let

$$\begin{aligned} p(x) &= x^4 - 1 \\ &= (x+1)(x-1)(x+i)(x-i). \end{aligned}$$

Let $m(x)$ be the minimal polynomial of T. Then m divides p, which implies m only has simple roots (no repeated roots), so T is diagonalisable, by 5.62.

(ii) Similar to the above, consider $p(x) = x^4 - x = x(x-1)(x+i)(x-i)$.

(iii) Consider

$$T = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Then we have that

$$T^2 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = T^4.$$

□

Exercise 6.9 ([Axl24] 5D Q2)**.** Suppose $T \in \mathscr{L}(V)$ has a diagonal matrix A with respect to some basis of V. Prove that if $\lambda \in \mathbf{F}$, then λ appears on the diagonal of A precisely $\dim E(\lambda, T)$ times.

Exercise 6.10 ([Axl24] 5D Q3)**.** Suppose V is finite-dimensional and $T \in \mathscr{L}(V)$. Prove that if the operator T is diagonalisable, then $V = \ker T \oplus \operatorname{im} T$.

Exercise 6.11 ([Axl24] 5D Q4)**.** Suppose V is finite-dimensional and $T \in \mathscr{L}(V)$. Prove that the following are equivalent.

(i) $V = \ker T \oplus \operatorname{im} T$.

(ii) $V = \ker T + \operatorname{im} T$.

(iii) $\ker T \cap \operatorname{im} T = \{\mathbf{0}\}$.

Exercise 6.12 ([Axl24] 5D Q5)**.** Suppose V is a finite-dimensional complex vector space and $T \in \mathscr{L}(V)$. Prove that T is diagonalisable if and only if

$$V = \ker(T - \lambda I) \oplus \operatorname{im}(T - \lambda I)$$

for every $\lambda \in \mathbb{C}$.

Exercise 6.13 ([Axl24] 5D Q9)**.** Suppose $R, T \in \mathscr{L}(\mathbf{F}^3)$ each have 2, 6, 7 as eigenvalues. Prove that there exists an invertible operator $S \in \mathscr{L}(\mathbf{F}^3)$ such that $R = S^{-1}TS$.

Exercise 6.14 ([Axl24] 5D Q14)**.** Suppose $\mathbf{F} = \mathbb{C}$, k is a positive integer, and $T \in \mathscr{L}(V)$ is invertible. Prove that T is diagonalisable if and only if T^k is diagonalisable.

Exercise 6.15 ([Axl24] 5D Q20)**.** Suppose V is finite-dimensional and $T \in \mathscr{L}(V)$. Prove that T is diagonalisable if and only if the dual operator T^* is diagonalisable.

Exercise 6.16 ([Axl24] 5E Q2)**.** Suppose $\mathscr{E} \subset \mathscr{L}(V)$ and every element of $\mathscr{E}$ is diagonalisable. Prove that there exists a basis of V with respect to which every element of $\mathscr{E}$ has a diagonal matrix if and only if every pair of elements of $\mathscr{E}$ commutes.

Exercise 6.17 ([Axl24] 5E Q3)**.** Suppose $S, T \in \mathscr{L}(V)$ are such that $ST = TS$. Suppose $p \in \mathbf{F}[z]$.

(i) Prove that $\ker p(S)$ is invariant under T.

(ii) Prove that $\operatorname{im} p(S)$ is invariant under T.

Exercise 6.18 ([Axl24] 5E Q4)**.** Prove or give a counterexample: If A is a diagonal matrix and B is an upper-triangular matrix of the same size as A, then A and B commute.

Solution. Counterexample:

$$\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$
$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix}$$

□

Exercise 6.19 ([Axl24] 5E Q5)**.** Prove that a pair of operators on a finite-dimensional vector space commute if and only if their dual operators commute.

Exercise 6.20 ([Axl24] 5E Q7)**.** Suppose V is a complex vector space, $S \in \mathscr{L}(V)$ is diagonalisable, and $T \in \mathscr{L}(V)$ commutes with S. Prove that there is a basis of V such that S has a diagonal matrix with respect to this basis and T has an upper-triangular matrix with respect to this basis.

7 Inner Product Spaces

7.1 Inner Products and Norms

Inner Products

Recall that we can define a dot product on the Euclidean space $\mathbb{R}^n$ as

$$x \cdot y = x_1 y_1 + \cdots + x_n y_n,$$

where $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$.

We now generalise this notion.

Definition 7.1 (Inner product space)**.** An ***inner product*** on V is a map $\langle \cdot, \cdot \rangle : V \times V \to \mathbf{F}$ such that for all $u, v, w \in V$, $\lambda \in \mathbf{F}$,

(i) $\langle v, v \rangle \geq 0$, where equality holds if and only if $v = \mathbf{0}$ (positive definite)

(ii) $\langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle$
$\langle \lambda u, v \rangle = \lambda \langle u, v \rangle$ (sesquilinear)

(iii) $\langle u, v \rangle = \overline{\langle v, u \rangle}$ (conjugate symmetric)

An ***inner product space*** $(V, \langle \cdot, \cdot \rangle)$ is a vector space V along with an inner product $\langle \cdot, \cdot \rangle$ on V.

Notation. If the inner product on V is clear from context, we omit it and simply denote the inner product space as V.

Remark. Every real number equals its complex conjugate. Thus if we are dealing with a real vector space, then in (iii) we can dispense with the complex conjugate, so $\langle u, v \rangle = \langle v, u \rangle$ for all $u, v \in V$.

Example.

- The *Euclidean inner product* on $\mathbf{F}^n$ is defined by

$$\langle (w_1,\dots,w_n),(z_1,\dots,z_n)\rangle = w_1\overline{z_1}+\cdots+w_n\overline{z_n}$$

  for all $(w_1,\dots,w_n),(z_1,\dots,z_n)\in\mathbf{F}^n$.

- An inner product can be defined on the vector space $\mathscr{C}([-1,1],\mathbb{R})$ by

$$\langle f,g\rangle = \int_{-1}^{1} fg$$

  for all $f,g\in\mathscr{C}([-1,1],\mathbb{R})$.

Lemma 7.2 (Basic properties of inner product)**.**

(i) *For each fixed $u\in V$, the function that sends $u\mapsto\langle u,v\rangle$ is a linear map from V to $\mathbf{F}$.*

(ii) $\langle 0,v\rangle = 0$ *for every $v\in V$.*

(iii) $\langle v,0\rangle = 0$ *for every $v\in V$.*

(iv) $\langle u,v+w\rangle = \langle u,v\rangle+\langle u,w\rangle$ *for all $u,v,w\in V$.*

(v) $\langle u,\lambda v\rangle = \overline{\lambda}\langle u,v\rangle$ *for all $\lambda\in\mathbf{F}$, $u,v\in V$.*

Proof.

(i) For $v\in V$, the linearity of $u\mapsto\langle u,v\rangle$ follows from the sesquilinearity of the inner product.

(ii) Every linear map takes $\mathbf{0}$ to 0. Thus (ii) follows from (i).

(iii) Let $v\in V$. By conjugate symmetry and (ii),

$$\langle v,0\rangle = \overline{\langle 0,v\rangle} = \overline{0} = 0.$$

(iv) Let $u,v,w\in V$. Then

$$\langle u,v+w\rangle = \overline{\langle v+w,u\rangle} = \overline{\langle v,u\rangle+\langle w,u\rangle} = \overline{\langle v,u\rangle}+\overline{\langle w,u\rangle} = \langle u,v\rangle+\langle u,w\rangle.$$

(v) Let $\lambda\in\mathbf{F}$, $u,v\in V$. Then

$$\langle u,\lambda v\rangle = \overline{\lambda v,u} = \overline{\lambda\langle v,u\rangle} = \overline{\lambda}\overline{\langle v,u\rangle} = \overline{\lambda}\langle u,v\rangle.$$

□

Norms

Each inner product determines a norm.

Definition 7.3 (Norm). For $v \in V$, the ***norm*** of v is

$$\|v\| := \sqrt{\langle v, v \rangle}.$$

Remark. Working with norms squared is usually easier than working directly with norms:

$$\|v\|^2 = \langle v, v \rangle .$$

Lemma 7.4 (Basic properties of norm). *Suppose $u, v \in V$.*

(i) $\|v\| \geq 0$, *where equality holds if and only if* $v = \mathbf{0}$. *(positive definiteness)*

(ii) $\|\lambda v\| = |\lambda| \|v\|$ *for all* $\lambda \in \mathbf{F}$. *(homogeneity)*

Proof.

(i) By positive definiteness of the inner product, $\langle v, v \rangle = 0$ if and only if $v = \mathbf{0}$. Take square roots to get $\|v\| = 0$.

(ii) Suppose $\lambda \in \mathbf{F}$. Then

$$\|\lambda v\|^2 = \langle \lambda v, \lambda v \rangle = \lambda \langle v, \lambda v \rangle = \lambda \overline{\lambda} \langle v, v \rangle = |\lambda|^2 \|v\|^2.$$

Taking square roots yields the desired equality.

□

Now we come to a crucial definition.

Definition 7.5 (Orthogonal vectors). We say $u, v \in V$ are ***orthogonal*** if $\langle u, v \rangle = 0$.

Lemma 7.6 (Orthogonality and $\mathbf{0}$).

(i) $\mathbf{0}$ *is orthogonal to every vector in* V.

(ii) $\mathbf{0}$ *is the only vector in* V *that is orthogonal to itself.*

Proof.

(i) Recall that $\langle \mathbf{0}, v \rangle = 0$ for every $v \in V$.

(ii) If $v \in V$ and $\langle v, v \rangle = 0$, then $v = \mathbf{0}$, by positive definiteness.

□

Lemma 7.7 (Pythagoras' theorem). *Suppose $u, v \in V$. If u and v are orthogonal, then*

$$\|u+v\|^2 = \|u\|^2 + \|v\|^2. \tag{7.1}$$

Proof. Suppose $\langle u, v \rangle = 0$. Then

$$\begin{aligned} \|u+v\|^2 &= \langle u+v, u+v \rangle \\ &= \langle u, u+v \rangle + \langle v, u+v \rangle \\ &= \langle u, u \rangle + \langle u, v \rangle + \langle v, u \rangle + \langle v, v \rangle \\ &= \|u\|^2 + 0 + \overline{0} + \|v\|^2 \\ &= \|u\|^2 + \|v\|^2 \end{aligned}$$

as desired. □

We now introdoce a process known as *orthogonal decomposition*. Suppose $u, v \in V$, $u \neq \mathbf{0}$. Then the ***orthogonal projection*** of v onto u is

$$\operatorname{proj}_u(v) := \frac{\langle v, u \rangle}{\langle u, u \rangle} u, \tag{7.2}$$

which is parallel to u. We check that $v - \operatorname{proj}_u(v)$ and u are orthogonal:

$$\begin{aligned} \langle v - \operatorname{proj}_u(v), u \rangle &= \langle v, u \rangle - \left\langle \frac{\langle v, u \rangle}{\langle u, u \rangle} u, u \right\rangle \\ &= \langle v, u \rangle - \frac{\langle v, u \rangle}{\langle u, u \rangle} \langle u, u \rangle = 0. \end{aligned}$$

Lemma 7.8 (Cauchy–Schwarz inequality). *Suppose $u, v \in V$. Then*

$$|\langle u, v \rangle| \leq \|u\| \|v\|, \tag{7.3}$$

where equality holds if and only if $u = \lambda v$ for some scalar λ.

Proof. If $u = \mathbf{0}$, then both sides of the desired inequality equal 0. Thus assume $u \neq \mathbf{0}$. Consider the orthogonal decomposition of v:

$$v = (v - \operatorname{proj}_u(v)) + \operatorname{proj}_u(v).$$

By the Pythagoras' theorem,

$$\|v\|^2 = \underbrace{\|v - \text{proj}_u(v)\|^2}_{\geq 0} + \|\text{proj}_u(v)\|^2,$$

so

$$\|v\| \geq \|\text{proj}_u(v)\| = \left|\frac{\langle v,u\rangle}{\langle u,u\rangle}\right| \|u\| = \frac{|\langle v,u\rangle|}{\|u\|}$$

and rearranging gives the desired inequality. Equality holds if and only if $v = \text{proj}_u(v)$, i.e.,

$$\frac{\langle v,u\rangle}{\langle u,u\rangle} u = v.$$

□

Lemma 7.9 (Triangle inequality). *Suppose $u, v \in V$. Then*

$$\|u+v\| \leq \|u\| + \|v\|, \tag{7.4}$$

where equality holds if and only if $u = \lambda v$ for some $\lambda \in \mathbb{R}_{\geq 0}$.

Proof. We have

$$\begin{aligned}
\|u+v\|^2 &= \langle u+v, u+v\rangle \\
&= \langle u,u\rangle + \langle v,v\rangle + \langle u,v\rangle + \langle v,u\rangle \\
&= \langle u,u\rangle + \langle v,v\rangle + \langle u,v\rangle + \overline{\langle u,v\rangle} \\
&= \|u\|^2 + \|v\|^2 + 2\,\text{Re}\,\langle u,v\rangle \\
&\leq \|u\|^2 + \|v\|^2 + 2|\langle u,v\rangle| && \text{(I)} \\
&\leq \|u\|^2 + \|v\|^2 + 2\|u\|\|v\| && \text{[by Cauchy–Schwarz inequality]} \quad \text{(II)} \\
&= (\|u\| + \|v\|)^2,
\end{aligned}$$

Taking square roots yields the desired inequality.

Equality holds if and only if equality holds in (I) and (II), i.e.,

$$\langle u,v\rangle = \|u\|\|v\|.$$

If $u = \lambda v$ for $\lambda \in \mathbb{R}_{\geq 0}$, then the above equation holds.

Conversely, suppose the above equation holds. Then equality in the Cauchy–Schwarz inequality implies that $u = \lambda v$ for some scalar λ. By the above equation, λ must be a non-negative real number, completing the proof. □

Corollary 7.10 (Reverse triangle inequality). *Suppose $u, v \in V$. Then*

$$|\|u\| - \|v\|| \leq \|u - v\|.$$

Proof. We have

$$\begin{aligned}
\|u-v\|^2 &= \langle u-v, u-v\rangle \\
&= \|u\|^2 + \|v\|^2 - (\langle u, v\rangle + \langle v, u\rangle) \\
&\geq \|u\|^2 + \|v\|^2 - 2\|u\|\|v\| \\
&= (\|u\| - \|v\|)^2.
\end{aligned}$$

Taking square roots yields the desired result. □

Lemma 7.11 (Parallelogram equality). *Suppose $u, v \in V$. Then*

$$\|u+v\|^2 + \|u-v\|^2 = 2\left(\|u\|^2 + \|v\|^2\right). \tag{7.5}$$

Proof. We have

$$\begin{aligned}
\|u+v\|^2 + \|u-v\|^2 &= \langle u+v, u+v\rangle + \langle u-v, u-v\rangle \\
&= \left(\|u\|^2 + \|v\|^2 + \langle u, v\rangle + \langle v, u\rangle\right) + \left(\|u\|^2 + \|v\|^2 - \langle u, v\rangle - \langle v, u\rangle\right) \\
&= 2\left(\|u\|^2 + \|v\|^2\right)
\end{aligned}$$

as desired. □

7.2 Orthonormal Bases

Orthonormal Bases

Definition 7.12 (Orthonormal basis)**.** We say $\{e_1,\dots,e_n\}$ is ***orthonormal*** if

(i) $\|e_i\| = 1$;

(ii) the vectors are pairwise orthogonal: $\langle e_i, e_j\rangle = \delta_{ij}$.

If additionally $\{e_1,\dots,e_n\}$ is a basis of V, then $\{e_1,\dots,e_n\}$ is a ***orthonormal basis*** of V.

The next result concerns the norm of an orthonormal linear combination.

Lemma 7.13. *Suppose $\{e_1,\dots,e_n\}$ is a orthonormal set of vectors in V. Then*

$$\|a_1e_1+\cdots+a_ne_n\|^2 = |a_1|^2+\cdots+|a_n|^2$$

for all $a_1,\dots,a_n \in \mathbf{F}$.

Proof. By the Pythagoras' theorem,

$$\begin{aligned}\|a_1e_1+\cdots+a_ne_n\|^2 &= \|a_1e_1\|^2+\cdots+\|a_ne_n\|^2\\ &= |a_1|^2\|e_1\|^2+\cdots+|a_n|^2\|e_n\|^2\\ &= |a_1|^2+\cdots+|a_n|^2\end{aligned}$$

since each $\|e_i\| = 1$. □

The result above has the following important corollary.

Corollary 7.14. *Every orthonormal set of vectors is linearly independent.*

Proof. Suppose $\{e_1,\dots,e_n\}$ is an orthonormal set of vectors in V. Suppose $a_1,\dots,a_n \in \mathbf{F}$ are such that

$$a_1e_1+\cdots+a_ne_n = \mathbf{0}.$$

By the previous result,

$$|a_1|^2+\cdots+|a_n|^2 = 0,$$

so $a_1 = \cdots = a_n = 0$. Hence $e_1,\dots,e_n$ are linearly independent. □

Corollary 7.15. *Suppose V is finite-dimensional. Then every orthonormal set of vectors in V of length $\dim V$ is an orthonormal basis of V.*

Proof. By 7.14, every orthonormal set of vectors in V is linearly independent. Thus by 3.32, every such set of length $\dim V$ is a basis. □

Now we come to an important inequality.

Lemma 7.16 (Bessel's inequality). *Suppose $\{e_1,\dots,e_n\}$ is an orthonormal set of vectors in V. If $v \in V$ then*

$$|\langle v,e_1\rangle|^2+\cdots+|\langle v,e_n\rangle|^2 \le \|v\|^2. \tag{7.6}$$

Proof. Let $v \in V$. For $i=1,\dots,n$, consider the orthogonal projection of v onto e_i:

$$\begin{aligned} v &= (v-\operatorname{proj}_{e_i}(v))+\operatorname{proj}_{e_i}(v)\\ &= \left(v-\frac{\langle v,e_i\rangle}{\langle e_i,e_i\rangle}e_i\right)+\frac{\langle v,e_i\rangle}{\langle e_i,e_i\rangle}e_i\\ &= (v-\langle v,e_i\rangle e_i)+\langle v,e_i\rangle e_i. \end{aligned}$$

Then by Pythagoras' theorem,

$$\begin{aligned} \|v\|^2 &= \|v-\langle v,e_i\rangle e_i\|^2+\|\langle v,e_i\rangle e_i\|\\ &= \|v-\langle v,e_i\rangle e_i\|^2+|\langle v,e_i\rangle|^2. \end{aligned}$$

Write

$$\begin{aligned} v &= \operatorname{proj}_{e_1}(v)+\cdots+\operatorname{proj}_{e_n}(v)+w\\ &= \langle v,e_1\rangle e_1+\cdots+\langle v,e_n\rangle e_n+w \end{aligned}$$

for some $w \in V$. Note that for $i=1,\dots,n$,

$$\begin{aligned} \langle v,e_i\rangle &= \langle\langle v,e_i\rangle e_i,e_i\rangle+\langle w,e_i\rangle\\ &= \langle v,e_i\rangle+\langle w,e_i\rangle \end{aligned}$$

which implies $\langle w,e_i\rangle = 0$, so w is orthogonal to $e_1,\dots,e_n$. Thus $e_1,\dots,e_n,w$ are pairwise orthogonal. By Pythagoras' theorem,

$$\begin{aligned} \|v\|^2 &= |\langle v,e_1\rangle|^2+\cdots+|\langle v,e_n\rangle|^2+\underbrace{\|w\|^2}_{\ge 0}\\ &\ge |\langle v,e_1\rangle|^2+\cdots+|\langle v,e_n\rangle|^2 \end{aligned}$$

as desired. Equality holds for orthonormal bases (as we will see later). □

The next result states that a vector can be expressed as a linear combination of an orthonormal basis. Usually we write $v=\sum_{i=1}^n a_i v_i$, but with orthonormal basis we can just take $a_i=\langle v,e_i\rangle$.

Lemma 7.17. *Suppose $\{e_1,\dots,e_n\}$ is an orthonormal basis of V, and $u,v\in V$. Then*

$$v=\langle v,e_1\rangle e_1+\cdots+\langle v,e_n\rangle e_n. \tag{7.7}$$

Proof. Since $\{e_1,\dots,e_n\}$ is a basis of V, there exist $a_1,\dots,a_n\in\mathbf{F}$ such that

$$v=a_1e_1+\cdots+a_ne_n.$$

Since $e_1,\dots,e_n$ are orthonormal, taking the inner product of both sides with e_i gives

$$\langle v,e_i\rangle=a_i\quad(i=1,\dots,n).$$

Hence we are done. □

Applying Pythagoras' theorem to (7.7), we obtain *Parseval's identity*:

$$\|v\|^2=|\langle v,e_1\rangle|^2+\cdots+|\langle v,e_n\rangle|^2. \tag{7.8}$$

Let $u,v\in V$. Taking the inner product of u on both sides of (7.7) gives

$$\begin{aligned}\langle u,v\rangle&=\langle u,\langle v,e_1\rangle e_1+\cdots+\langle v,e_n\rangle e_n\rangle\\&=\langle u,\langle v,e_1\rangle e_1\rangle+\cdot+\langle u,\langle v,e_n\rangle e_n\rangle\\&=\overline{\langle v,e_1\rangle}\langle u,e_1\rangle+\cdots+\overline{\langle v,e_n\rangle}\langle u,e_n\rangle\end{aligned}$$

that is,

$$\langle u,v\rangle=\langle u,e_1\rangle\overline{\langle v,e_1\rangle}+\cdots+\langle u,e_n\rangle\overline{\langle v,e_n\rangle}. \tag{7.9}$$

Gram–Schmidt Procedure

The *Gram–Schmidt procedure* is a method for constructing orthonormal basis; it turns a linearly independent set into an orthonormal set with the same span as the original set. It guarantees the existence of orthonormal bases.

Theorem 7.18 (Gram–Schmidt procedure)**.** *Suppose $v_1,\dots,v_n$ are linearly independent in V. Define $u_1=v_1$, and for $i=2,\dots,n$,*

$$u_i=v_i-\operatorname{proj}_{u_1}(v_i)-\cdots-\operatorname{proj}_{u_{i-1}}(v_i).$$

Let $e_i=\dfrac{u_i}{\|u_i\|}$. Then $\{e_1,\dots,e_n\}$ is orthonormal in V, and that

$$\operatorname{span}(v_1,\dots,v_i)=\operatorname{span}(e_1,\dots,e_i)\quad(i=1,\dots,n).$$

Proof. Induct on i.

For the base case $i = 1$, since $e_1 = \frac{u_1}{\|u_1\|}$ we have $\|e_1\| = 1$, and $\operatorname{span}(v_1) = \operatorname{span}(e_1)$ because e_1 is a non-zero multiple of v_1.

Suppose the desired result holds for $i-1$; that is, the set $\{e_1, \ldots, e_{i-1}\}$ generated by the above procedure is an orthonormal set, and

$$\operatorname{span}(v_1, \ldots, v_{i-1}) = \operatorname{span}(e_1, \ldots, e_{i-1}). \tag{I}$$

Since $v_1, \ldots, v_n$ are linearly independent, we have $v_i \notin \operatorname{span}(v_1, \ldots, v_{i-1})$. Thus $v_i \notin \operatorname{span}(e_1, \ldots, e_{i-1}) = \operatorname{span}(u_1, \ldots, u_{i-1})$, which implies that $u_i \neq \mathbf{0}$ (so we are not dividing by 0); thus $\|e_i\| = 1$.

We now check that $\{e_1, \ldots, e_i\}$ is an orthonormal set. For $j \in \{1, \ldots, i-1\}$,

$$\begin{aligned}
\langle e_i, e_j \rangle &= \left\langle \frac{u_i}{\|u_i\|}, \frac{u_j}{\|u_j\|} \right\rangle \\
&= \frac{1}{\|u_i\|\|u_j\|} \langle u_i, u_j \rangle \\
&= \frac{1}{\|u_i\|\|u_j\|} \left\langle v_i - \operatorname{proj}_{u_1}(v_i) - \cdots - \operatorname{proj}_{u_j}(v_i) - \cdots - \operatorname{proj}_{u_{i-1}}(v_i), u_j \right\rangle \\
&= \frac{1}{\|u_i\|\|u_j\|} \left\langle v_i - \frac{\langle v_i, u_1 \rangle}{\langle u_1, u_1 \rangle} u_1 - \cdots - \frac{\langle v_i, u_j \rangle}{\langle u_j, u_j \rangle} u_j - \cdots - \frac{\langle v_i, u_{i-1} \rangle}{\langle u_{i-1}, u_{i-1} \rangle} u_{i-1}, u_j \right\rangle \\
&= \frac{1}{\|u_i\|\|u_j\|} \left(\langle v_i, u_j \rangle - \left\langle \frac{\langle v_i, u_j \rangle}{\langle u_j, u_j \rangle} u_j, u_j \right\rangle \right) \\
&= \frac{1}{\|u_i\|\|u_j\|} \left(\langle v_i, u_j \rangle - \langle v_i, u_j \rangle \right) = 0
\end{aligned}$$

so e_i is orthogonal to $e_1, \ldots, e_{i-1}$. Hence $\{e_1, \ldots, e_i\}$ is an orthonormal set of vectors.

From the definition of e_i, we see that $v_i \in \operatorname{span}(e_1, \ldots, e_i)$. Combining this information with (I) shows that

$$\operatorname{span}(v_1, \ldots, v_i) \subset \operatorname{span}(e_1, \ldots, e_i).$$

Both sets above are linearly independent (the v's by hypothesis, and the e's by orthonormality and 7.14). Thus both subspaces above have dimension i, and hence they are equal, completing the induction step and thus completing the proof. □

Now we can answer the question about the *existence* of orthonormal bases.

Corollary 7.19. *Every finite-dimensional inner product space has an orthonormal basis.*

Proof. Suppose V is finite-dimensional. Choose a basis of V.

Apply the Gram–Schmidt procedure (7.18) to it, producing an orthonormal set of length $\dim V$. By 7.15, this orthonormal set is an orthonormal basis of V. □

Sometimes we need to know not only that an orthonormal basis exists, but also that every orthonormal set can be extended to an orthonormal basis. In the next corollary, the Gram–Schmidt procedure shows that such an extension is always possible.

Corollary 7.20. *Suppose V is finite-dimensional. Then every orthonormal set of vectors in V can be extended to an orthonormal basis of V.*

Proof. Suppose $\{e_1,\dots,e_m\}$ is an orthonormal set of vectors in V. By 7.14, $\{e_1,\dots,e_m\}$ is linearly independent, and thus can be extended to a basis $\{e_1,\dots,e_m,v_1,\dots,v_n\}$ of V.

Now apply the Gram–Schmidt procedure to $\{e_1,\dots,e_m,v_1,\dots,v_n\}$, producing an orthonormal set

$$\{e_1,\dots,e_m,u_1,\dots,u_n\}$$

where the first m vectors are unchanged because they are already orthonormal. The set above is an orthonormal basis of V by 7.15. □

The next result shows that the condition for an operator to have an upper-triangular matrix with respect to some *orthonormal basis* is the same as the condition to have an upper-triangular matrix with respect to an *arbitrary basis* (recall 6.24).

Proposition 7.21. *Suppose V is finite-dimensional, $T \in \mathscr{L}(V)$. Then T has an upper-triangular matrix with respect to some orthonormal basis of V if and only if the minimal polynomial of T equals*

$$(z-\lambda_1)\cdots(z-\lambda_n)$$

for some $\lambda_1,\dots,\lambda_n \in \mathbf{F}$.

Proof. Suppose T has an upper-triangular matrix with respect to some basis $\{v_1,\dots,v_n\}$ of V. By 6.21, $\operatorname{span}(v_1,\dots,v_i)$ is invariant under T for each $i = 1,\dots,n$.

Apply the Gram–Schmidt procedure to $\{v_1,\dots,v_n\}$, producing an orthonormal basis $\{e_1,\dots,e_n\}$ of V. Since

$$\operatorname{span}(e_1,\dots,e_i) = \operatorname{span}(v_1,\dots,v_i),$$

we conclude that $\operatorname{span}(e_1,\dots,e_i)$ is invariant under T for each $i = 1,\dots,n$. Thus by 6.21, T has an upper-triangular matrix with respect to the orthonormal basis $\{e_1,\dots,e_n\}$.

By 6.24, the minimal polynomial of T equals $(z-\lambda_1)\cdots(z-\lambda_n)$ for some $\lambda_1,\dots,\lambda_n \in \mathbf{F}$. □

Theorem 7.22 (Schur's theorem). *Every operator on a finite-dimensional complex inner product space has an upper-triangular matrix with respect to some orthonormal basis.*

Proof. The desired result follows from the second version of the fundamental theorem of algebra (5.11) and 7.21. □

Linear Functionals on Inner Product Spaces

Theorem 7.23 (Riesz representation theorem). *Suppose V is finite-dimensional, and ϕ is a linear functional on V. Then for every $u \in V$, there exists a unique $v \in V$ such that*

$$\phi(u) = \langle u, v \rangle .$$

Proof.

Existence Pick an orthonormal basis $\{e_1, \ldots, e_n\}$ of V. Let $u \in V$. By 7.17,

$$u = \langle u, e_1 \rangle e_1 + \cdots + \langle u, e_n \rangle e_n.$$

Applying ϕ on u gives

$$\begin{aligned}
\phi(u) &= \phi\left(\langle u, e_1 \rangle e_1 + \cdots + \langle u, e_n \rangle e_n\right) \\
&= \langle u, e_1 \rangle \phi(e_1) + \cdots + \langle u, e_n \rangle \phi(e_n) \\
&= \left\langle u, \overline{\phi(e_1)} e_1 \right\rangle + \cdots + \left\langle u, \overline{\phi(e_n)} e_n \right\rangle \\
&= \left\langle u, \overline{\phi(e_1)} e_1 + \cdots + \overline{\phi(e_n)} e_n \right\rangle .
\end{aligned}$$

Pick

$$v = \overline{\phi(e_1)} e_1 + \cdots + \overline{\phi(e_n)} e_n.$$

Then we have $\phi(u) = \langle u, v \rangle$ for every $u \in V$, as desired.

Uniqueness Suppose $v, v' \in V$ satisfy

$$\phi(u) = \langle u, v \rangle = \left\langle u, v' \right\rangle$$

for every $u \in V$. Then

$$0 = \langle u, v \rangle - \left\langle u, v' \right\rangle = \left\langle u, v - v' \right\rangle$$

for every $u \in V$. Taking $u = v - v'$ shows that $v - v' = \mathbf{0}$, so $v = v'$. □

7.3 Orthogonal Complements and Minimisation Problems

Orthogonal Complements

Definition 7.24 (Orthogonal complement)**.** The ***orthogonal complement*** of $U \subset V$ is

$$U^{\perp} := \{v \in V \mid \langle u, v\rangle = 0, \forall u \in U\}.$$

That is, $U^{\perp}$ is the set of vectors in V that are orthogonal to every vector in U.

We check that if $U \subset V$, then $U^{\perp} \leq V$:

(i) $\langle u, \mathbf{0}\rangle = 0$ for every $u \in U$, so $\mathbf{0} \in U^{\perp}$.

(ii) Let $v, w \in U^{\perp}$. For every $u \in U$,

$$\langle u, v+w\rangle = \langle u, v\rangle + \langle u, w\rangle = 0+0 = 0 \implies v+w \in U^{\perp}$$

so $U^{\perp}$ is closed under addition.

(iii) Let $v \in U^{\perp}$, $\lambda \in \mathbf{F}$. For every $u \in U$,

$$\langle u, \lambda v\rangle = \overline{\lambda}\,\langle u, v\rangle = \overline{\lambda} \cdot 0 = 0 \implies \lambda v \in U^{\perp}$$

so $U^{\perp}$ is closed under scalar multiplication.

Example.

- Let U be a plane in $\mathbb{R}^3$ containing the origin. Then $U^{\perp}$ is the line containing the origin that is perpendicular to U.
- Let U be a line in $\mathbb{R}^3$ containing the origin. Then $U^{\perp}$ is the plane containing the origin that is perpendicular to U.

We begin with some straightforward consequences of the definition.

Lemma 7.25 (Properties of orthogonal complement)**.**

(i) $\{\mathbf{0}\}^{\perp} = V$, $V^{\perp} = \{\mathbf{0}\}$.

(ii) If $U \subset V$, *then* $U \cap U^{\perp} \subset \{\mathbf{0}\}$.

(iii) If $G \subset H \subset V$, *then* $H^{\perp} \subset G^{\perp}$.

Proof.

(i)
$$v \in \{\mathbf{0}\}^\perp \iff \langle \mathbf{0}, v\rangle = 0 \iff v \in V$$
$$v \in V^\perp \iff \langle v, v\rangle = 0 \iff v = \mathbf{0}$$

(ii) Suppose $U \subset V$. Let $u \in U \cap U^\perp$, then $\langle u, u\rangle = 0$ so $u = \mathbf{0}$. Hence $U \cap U^\perp \subset \{\mathbf{0}\}$.

(iii) Suppose $G \subset H \subset V$. Let $v \in H^\perp$, then

$$\langle u, v\rangle = 0 \quad (\forall u \in H)$$

which implies

$$\langle u, v\rangle = 0 \quad (\forall u \in G).$$

Hence $v \in G^\perp$, so $H^\perp \subset G^\perp$.

□

The next result shows that every *finite-dimensional* subspace of V leads to a natural direct sum decomposition of V.

Lemma 7.26. *Suppose $U \leq V$ is finite-dimensional. Then*

$$V = U \oplus U^\perp.$$

Proof. We first show that $V = U + U^\perp$. Let $v \in V$, pick an orthonormal basis $\{e_1, \ldots, e_m\}$ of U. We can write

$$v = \underbrace{\langle v, e_1\rangle e_1 + \cdots + \langle v, e_m\rangle e_m}_{u} + \underbrace{v - \langle v, e_1\rangle e_1 - \cdots - \langle v, e_m\rangle e_m}_{w}. \tag{I}$$

We are left to check that $u \in U$ and $w \in U^\perp$.

- Since each $u_i \in U$, we see that $u \in U$.
- Since $\{e_1, \ldots, e_m\}$ is an orthonormal set of vectors, for each $i = 1, \ldots, m$,

$$\begin{aligned}\langle w, e_i\rangle &= \langle v - \langle v, e_1\rangle e_1 - \cdots - \langle v, e_m\rangle e_m, e_i\rangle \\ &= \langle v, e_i\rangle - \langle v, e_i\rangle = 0.\end{aligned}$$

Thus w is orthogonal to every vector in $\operatorname{span}(e_1, \ldots, e_m)$, which shows that $w \in U^\perp$.

Since $U \cap U^\perp = \{\mathbf{0}\}$, by 3.13, $U + U^\perp$ is a direct sum. □

Corollary 7.27. *Suppose V is finite-dimensional and $U \leq V$. Then*

$$\dim U^\perp = \dim V - \dim U.$$

Proof. Apply 4.50 for the dimension of a direct sum. □

Corollary 7.28. *Suppose $U \leq V$ is finite-dimensional. Then*

$$U = (U^\perp)^\perp.$$

Proof.

$\boxed{\subset}$ Let $u \in U$. Then $\langle u, w\rangle = 0$ for every $w \in U^\perp$ (by definition of $U^\perp$). Since u is orthogonal to every vector in $U^\perp$, we have $u \in (U^\perp)^\perp$.

Hence $U \subset (U^\perp)^\perp$.

$\boxed{\supset}$ Let $v \in (U^\perp)^\perp$. Since $U + U^\perp$ is a direct sum, $v = u + w$ for some $u \in U$, $w \in U^\perp$.

Then $v - u = w \in U^\perp$. Since $v \in (U^\perp)^\perp$ and $u \in (U^\perp)^\perp$ (as $U \subset (U^\perp)^\perp$), we have $v - u \in (U^\perp)^\perp$.

Thus $v - u \in U^\perp \cap (U^\perp)^\perp$, which implies that $v - u = \mathbf{0}$, so $v = u$, and thus $v \in U$.

Hence $(U^\perp)^\perp \subset U$. □

Suppose U is a subspace of V and we want to show that $U = V$. Sometimes the easiest way to do so is to show that the only vector orthogonal to U is $\mathbf{0}$, and then use the result below.

Corollary 7.29. *Suppose $U \leq V$ is finite-dimensional. Then*

$$U^\perp = \{\mathbf{0}\} \iff U = V.$$

Proof.

$\boxed{\Longrightarrow}$ Suppose $U^\perp = \{\mathbf{0}\}$. Then $U = (U^\perp)^\perp = \{\mathbf{0}\}^\perp = V$, as desired.

$\boxed{\Longleftarrow}$ If $U = V$, then $U^\perp = V^\perp = \{\mathbf{0}\}$. □

We now define an operator P_U for each finite-dimensional subspace U of V.

Definition 7.30 (Orthogonal projection). Suppose $U \leq V$ is finite-dimensional. The ***orthogonal projection*** is the operator $P_U \in \mathscr{L}(V)$ defined as follows: For each $v \in V$, write $v = u + w$ for some $u \in U$, $w \in U^\perp$. Then let $P_U v = u$.

Remark. The direct sum decomposition $V = U \oplus U^\perp$ shows that each $v \in V$ can be uniquely written in the form $v = u + w$ with $u \in U$, $w \in U^\perp$. Thus $P_U v$ is well defined.

Suppose $u \in V$ with $u \neq \mathbf{0}$ and $U = \operatorname{span}(u)$. If $v \in V$ then

$$v = \frac{\langle v,u\rangle}{\|u\|^2}u + \left(v - \frac{\langle v,u\rangle}{\|u\|^2}u\right).$$

Then this implies that

$$P_U v := \frac{\langle v,u\rangle}{\|u\|^2}u.$$

We now check that $P_U \in \mathscr{L}(V)$.

(i) Let $v_1, v_2 \in V$. Write

$$v_1 = u_1 + w_1, \quad v_2 = u_2 + w_2$$

for some $u_1, u_2 \in U$, $w_1, w_2 \in U^\perp$. Thus $P_U v_1 = u_1$ and $P_U v_2 = u_2$. Since

$$v_1 + v_2 = (\underbrace{u_1 + u_2}_{\in U}) + (\underbrace{w_1 + w_2}_{\in U^\perp}),$$

we have

$$P_U(v_1 + v_2) = u_1 + u_2 = P_U v_1 + P_U v_2.$$

(ii) Let $\lambda \in \mathbf{F}$, $v \in V$. Write $v = u + w$, where $u \in U$, $w \in U^\perp$. Then

$$\lambda v = \underbrace{\lambda u}_{\in U} + \underbrace{\lambda w}_{\in U^\perp},$$

so

$$P_U(\lambda v) = \lambda u = \lambda P_U v.$$

Lemma 7.31 (Properties of orthogonal projection). *Suppose $U \leq V$ is finite-dimensional.*

(i) $P_U u = u$ *for every* $u \in U$, $P_U w = \mathbf{0}$ *for every* $w \in U^\perp$.

(ii) $\operatorname{im} P_U = U$, $\ker P_U = U^\perp$.

(iii) $v - P_U v \in U^\perp$ *for every* $v \in V$.

(iv) ${P_U}^2 = P_U$.

(v) $\|P_U v\| \leq \|v\|$ *for every* $v \in V$.

(vi) If $\{e_1, \ldots, e_n\}$ *is an orthonormal basis of* U, *and* $v \in V$, *then*

$$P_U v = \langle v, e_1\rangle e_1 + \cdots + \langle v, e_n\rangle e_n.$$

Proof.

(i) Let $u \in U$. We can write $u = u + \mathbf{0}$, where $u \in U$, $\mathbf{0} \in U^\perp$. Thus $P_U u = u$.

Let $w \in U^\perp$. We can write $w = \mathbf{0} + w$, where $\mathbf{0} \in U$, $w \in U^\perp$. Thus $P_U w = \mathbf{0}$.

(ii) The definition of P_U implies that $\operatorname{im} P_U \subset U$. Furthermore, (i) implies that $U \subset \operatorname{im} P_U$. Hence $\operatorname{im} P_U = U$.

The inclusion $U^\perp \subset \ker P_U$ follows from (i). To prove the inclusion in the other direction, if $v \in \ker P_U$, then the decomposition given by 7.26 must be $v = \mathbf{0} + v$, where $\mathbf{0} \in U$ and $v \in U^\perp$. Thus $\ker P_U \subset U^\perp$.

(iii) If $v \in V$ and $v = u + w$ with $u \in U$, $w \in U^\perp$, then

$$v - P_U v = v - u = w \in U^\perp.$$

(iv) If $v \in V$ and $v = u + w$ with $u \in U$, $w \in U^\perp$, then

$$(P_U{}^2)v = P_U(P_U v) = P_U u = u = P_U v.$$

(v) If $v \in V$ and $v = u + w$ with $u \in U$, $w \in U^\perp$, then

$$\|P_U v\|^2 \leq \|u\|^2 \leq \|u\|^2 + \|w\|^2 = \|v\|^2,$$

where the last equality comes from the Pythagorean theorem.

(vi) The formula for $P_U v$ follows from equation (I) in the proof of 7.26.

□

Minimisation Problems

Given a subspace U of V and a point $v \in V$, we want to find a point $u \in U$ that minimises $\|v - u\|$. The next result shows that $u = P_U v$ is the unique solution of this minimisation problem.

Theorem 7.32 (Minimising distance to a subspace). *Suppose $U \leq V$ is finite-dimensional. Fix $v \in V$. Then for all $u \in U$,*

$$\|v - P_U v\| \leq \|v - u\|, \tag{7.10}$$

where equality holds if and only if $u = P_U v$.

Proof. We have

$$\begin{aligned} \|v - P_U v\|^2 &\leq \|v - P_U v\|^2 + \|P_U v - u\|^2 && [\because \|P_U v - u\|^2 \geq 0] \\ &= \|(v - P_U v) + (P_U v - u)\|^2 && [\text{by Pythagoras' theorem}] \\ &= \|v - u\|^2. \end{aligned}$$

Taking square roots gives the desired inequality. Equality holds if and only if $\|P_U v - u\| = 0$, which holds if and only if $u = P_U v$. □

insert figure

Pseudoinverse

Suppose $T \in \mathscr{L}(V,W)$ and $w \in W$. Consider the problem of finding $v \in V$ such that

$$Tv = w.$$

If T is invertible, then evidently we are done. The pseudoinverse will provide the tool to solve the equation above as well as possible, even when T is not invertible.

We need the next result to define the pseudoinverse; it states that we can restrict a linear map to obtain a bijective map.

Lemma 7.33. *Suppose V is finite-dimensional, $T \in \mathscr{L}(V)$. Then $T|_{(\ker T)^\perp}$ is an bijective map from $(\ker T)^\perp$ to $\operatorname{im} T$.*

Proof. To prove bijectivity, we need to show injectivity and surjectivity.

Injectivity Let $v \in (\ker T)^\perp$ be such that $v \in \ker T|_{(\ker T)^\perp}$. Then

$$\begin{aligned} T|_{(\ker T)^\perp} v = \mathbf{0} &\implies Tv = \mathbf{0} \\ &\implies v \in (\ker T) \cap (\ker T)^\perp \\ &\implies v = \mathbf{0} && [\text{by 7.25}] \end{aligned}$$

Hence $\ker T|_{(\ker T)^\perp} = \{\mathbf{0}\}$, so $T|_{(\ker T)^\perp}$ is injective.

Surjectivity Clearly $\operatorname{im} T|_{(\ker T)^\perp} \subset \operatorname{im} T$. To prove the inclusion in the other direction, let $w \in \operatorname{im} T$, so there exists $v \in V$ such that $w = Tv$.

By 7.26, $V = \ker T \oplus (\ker T)^\perp$. Thus $v = u + x$ for some $u \in \ker T$, $x \in (\ker T)^\perp$. Now

$$T|_{(\ker T)^\perp} x = Tx = Tv - Tu = w - \mathbf{0} = w,$$

which shows that $w \in \operatorname{im} T|_{(\ker T)^\perp}$, so $\operatorname{im} T \subset \operatorname{im} T|_{(\ker T)^\perp}$. Hence $\operatorname{im} T|_{(\ker T)^\perp} = \operatorname{im} T$.

□

Now we can define the *pseudoinverse* of a linear map T. In the next definition (and from now on), we can think of $T|_{(\ker T)^\perp}$ as an invertible linear map from $(\ker T)^\perp$ to $\operatorname{im} T$, as is justified by the result above.

Definition 7.34 (Pseudoinverse)**.** Suppose V is finite-dimensional, $T \in \mathscr{L}(V)$. The ***pseudoinverse*** (or *Moore–Penrose inverse*) $T^+ \in \mathscr{L}(W,V)$ of T is defined by

$$T^+ w := \left(T|_{(\ker T)^\perp}\right)^{-1} P_{\operatorname{im} T} w \quad (w \in W).$$

The pseudoinverse behaves much like an inverse, as we will see.

Lemma 7.35 (Properties of pseudoinverse)**.** *Suppose V is finite-dimensional, and $T \in \mathscr{L}(V)$.*

(i) If T is invertible, then $T^+ = T^{-1}$.

(ii) $TT^+ = P_{\operatorname{im} T}$ (orthogonal projection of W onto $\operatorname{im} T$).

(iii) $T^+T = P_{(\ker T)^\perp}$ (orthogonal projection of V onto $(\ker T)^\perp$).

Proof.

(i) Suppose T is invertible. Then injectivity implies $(\ker T)^\perp = V$, and surjectivity implies $\operatorname{im} T = W$.

Thus $T|_{(\ker T)^\perp} = T$, and $P_{\operatorname{im} T} = I_W$. Hence $T^+ = T$.

(ii) Let $w \in \operatorname{im} T$. Thus

$$TT^+ w = T\left(T|_{(\ker T)^\perp}\right)^{-1} w = w = P_{\operatorname{im} T} w.$$

Let $w \in (\operatorname{im} T)^\perp$, then $T^+ w = \mathbf{0}$. Hence $TT^+ w = \mathbf{0} = P_{\operatorname{im} T} w$.

Thus TT^+ and $P_{\operatorname{im} T}$ agree on $\operatorname{im} T$ and on $(\operatorname{im} T)^\perp$. Hence these two linear maps are equal (by 7.26).

(iii) Let $v \in (\ker T)^\perp$. Since $Tv \in \operatorname{im} T$, the definition of T^+ shows that

$$T^+(Tv) = \left(T|_{(\ker T)^\perp}\right)^{-1} (Tv) = v = P_{(\ker T)^\perp} v.$$

Thus T^+T and $P_{(\ker T)^\perp}$ agree on $(\ker T)^\perp$ and on $\ker T$. Hence these two linear maps are equal (by 7.26).

□

For $T \in \mathscr{L}(V,W)$ and $w \in W$, we now return to the problem of finding $v \in V$ that solves the equation

$$Tv = w.$$

As we noted earlier, if T is invertible, then $v = T^{-1}w$ is the unique solution, but if T is not invertible, then T^{-1} is not defined. However, the pseudoinverse T^+ is defined.

In the next result, (i) shows that taking $v = T^+w$ makes Tv as close to w as possible; thus the pseudoinverse provides a *best fit* to the equation above. (ii) shows that among all vectors $v \in V$ that make Tv as close as possible to w, the vector T^+w has the smallest norm.

Theorem 7.36 (Pseudoinverse provides best approximate solution or best solution). *Suppose V is finite-dimensional, $T \in \mathscr{L}(V,W)$, and $w \in W$.*

(i) If $v \in V$, then

$$\left\|T\left(T^+w\right) - w\right\| \le \|Tv - w\|, \tag{7.11}$$

where equality holds if and only if $v \in T^+w + \operatorname{ker} T$.

(ii) If $v \in T^+w + \operatorname{ker} T$, then

$$\left\|T^+w\right\| \le \|v\|, \tag{7.12}$$

where equality holds if and only if $v = T^+w$.

Proof.

(i) Let $v \in V$. Then

$$Tv - w = (Tv - TT^+w) + (TT^+w - w).$$

The first term in parentheses above is in $\operatorname{im} T$. Since the operator TT^+ is the orthogonal projection of W onto $\operatorname{im} T$ (by 7.35), the second term in parentheses above is in $(\operatorname{im} T)^\perp$ (by 7.31).

Thus the Pythagorean theorem implies the desired inequality that the norm of the second term in parentheses above is less than or equal to $\|Tv - w\|$, where equality holds if and only if the first term in parentheses above equals 0. Hence equality holds if and only if $v - T^+w + \operatorname{ker} T$, which is equivalent to the statement that $v \in T^+w + \operatorname{ker} T$.

(ii) Let $v \in T^+w + \operatorname{ker} T$. Then $v - T^+w \in \operatorname{ker} T$. Now

$$v = \left(v - T^+w\right) + T^+w.$$

The definition of T^+ implies that $T^+w \in (\operatorname{ker} T)^\perp$. Thus the Pythagorean theorem implies that $\|T^+w\| \le \|v\|$, where equality holds if and only if $v = T^+w$.

□

Exercises

Exercise 7.1 ([Axl24] 6A Q1)**.** Show that if $v_1, \dots, v_m \in V$, then

$$\sum_{j=1}^{m} \sum_{k=1}^{m} \langle v_j, v_k \rangle \geq 0.$$

Solution.

$$\sum_{j=1}^{m} \left(\sum_{k=1}^{m} \langle v_j, v_k \rangle \right) = \sum_{j=1}^{m} \left\langle v_j, \sum_{k=1}^{m} v_k \right\rangle = \left\langle \sum_{j=1}^{m} v_j, \sum_{k=1}^{m} v_k \right\rangle = \left\| \sum_{k=1}^{m} v_k \right\|^2 \geq 0.$$

□

Exercise 7.2 ([Axl24] 6A Q2)**.** Suppose $S \in \mathscr{L}(V)$. Define $\langle \cdot, \cdot \rangle_1$ by

$$\langle u, v \rangle_1 = \langle Su, Sv \rangle$$

for all $u, v \in V$. Show that $\langle \cdot, \cdot \rangle_1$ is an inner product on V if and only if S is injective.

Solution.

$\boxed{\Longrightarrow}$ Suppose $\langle \cdot, \cdot \rangle_1$ is an inner product on V. Let $u \in \ker S$, then

$$Su = \mathbf{0} \implies \langle Su, Su \rangle = \langle u, u \rangle_1 = 0 \implies u = \mathbf{0}.$$

Hence $\ker S = \{\mathbf{0}\}$, so S is injective.

$\boxed{\Longleftarrow}$ Check conditions for inner product:

(i) $\langle u, u \rangle_1 = \langle Su, Su \rangle \geq 0$, and $\langle u, u \rangle_1 = 0 \iff \langle Su, Su \rangle = 0 \iff Su = \mathbf{0} \iff u = \mathbf{0}$ by injectivity of S.

(ii)

(iii)

□

Exercise 7.3 ([Axl24] 6A Q4, modified)**.** Suppose $T \in \mathscr{L}(V)$ is a *contraction*; that is, $\|Tv\| \leq \|v\|$ for every $v \in V$. Prove that if $|\lambda| > 1$, then $T - \lambda I$ is injective.

Solution. Let $v \in \ker(T - \lambda I)$, then $Tv = \lambda v$, so

$$\begin{aligned} & \|Tv\| = \|\lambda v\| = |\lambda|\|v\| \\ \implies & |\lambda|\|v\| \leq \|v\| \\ \implies & (\underbrace{|\lambda| - 1}_{>0})\|v\| \leq 0 \\ \implies & \|v\| \leq 0 \end{aligned}$$

and hence $v = \mathbf{0}$. □

Exercise 7.4 ([Axl24] 6A Q5)**.** Suppose V is a real inner product space.

(i) Show that $\langle u+v, u-v\rangle = \|u\|^2 - \|v\|^2$ for every $u, v \in V$.

(ii) Show that if $u, v \in V$ have the same norm, then $u+v$ is orthogonal to $u-v$.

(iii) Use (ii) to show that the diagonals of a rhombus are perpendicular to each other.

Solution.

(i) We have that

$$\begin{aligned} \langle u+v, u-v\rangle &= \langle u, v\rangle - \langle v, v\rangle - \langle u, v\rangle + \langle v, u\rangle \\ &= \|u\|^2 - \|v\|^2 \end{aligned}$$

(ii) We know $\|u\| = \|v\|$, then

$$\langle u+v, u-v\rangle = \|u\|^2 - \|v\|^2 = 0$$

which shows that they are orthogonal.

(iii)

□

Exercise 7.5 ([Axl24] 6A Q6)**.** Suppose $u, v \in V$. Prove that $\langle u, v\rangle = 0 \iff \|u\| \leq \|u + av\|$ for all $a \in \mathbf{F}$.

Solution.

$\boxed{\Longrightarrow}$ We have

$$\begin{aligned}\|u+av\|^2 &= \langle u+av, u+av\rangle \\ &= \langle u,u\rangle + a\underbrace{\langle v,u\rangle}_{0} + \overline{a}\underbrace{\langle u,v\rangle}_{0} + |a|^2\langle v,v\rangle \\ &= \|u\|^2 + |a|^2\|v\|^2 \geq \|u\|^2.\end{aligned}$$

$\boxed{\Longleftarrow}$ If $v \neq \mathbf{0}$, then it is trivial. Assume $v \neq \mathbf{0}$. Let $a = \frac{\langle u,v\rangle}{\|v\|^2}$. Then we have

$$\begin{aligned}\left\|u - \frac{\langle u,v\rangle}{\|v\|^2}v\right\|^2 &= \left\langle u - \frac{\langle u,v\rangle}{\|v\|^2}v, u - \frac{\langle u,v\rangle}{\|v\|^2}v\right\rangle \\ &= \|u\|^2 - \frac{\overline{\langle u,v\rangle}}{\|v\|^2}\langle u,v\rangle - \frac{\langle u,v\rangle}{\|v\|^2}\langle v,u\rangle + \left|\frac{\langle u,v\rangle}{\|v\|^2}\right|^2\|v\|^2 \\ &= \|u\|^2 - 2\frac{|\langle u,v\rangle|^2}{\|v\|^2} + \frac{|\langle u,v\rangle|^2}{\|v\|^2} \\ &= \|u\|^2 - \frac{|\langle u,v\rangle|^2}{\|v\|^2} \geq \|u\|^2.\end{aligned}$$

□

Exercise 7.6 ([Axl24] 6A Q9)**.** Suppose $u, v \in V$ and $\|u\| = \|v\| = 1$ and $\langle u,v\rangle = 1$. Prove that $u = v$.

Solution. Cauchy–Schwarz inequality. □

Exercise 7.7 ([Axl24] 6A Q14)**.** Suppose $v \in V \setminus \{\mathbf{0}\}$. Prove that $v/\|v\|$ is the unique closest element on the unit sphere of V to v. More precisely, prove that if $u \in V$ and $\|u\| = 1$, then

$$\left\|v - \frac{v}{\|v\|}\right\| \leq \|v-u\|,$$

where equality holds if and only if $u = v/\|v\|$.

Solution. We have

$$\begin{aligned}\left\|v - \frac{v}{\|v\|}\right\| &= \left\|\left(1 - \frac{1}{\|v\|}\right)v\right\| \\ &= \left|1 - \frac{1}{\|v\|}\right|\|v\| \\ &= \left|\frac{\|v\|-1}{\|v\|}\right|\|v\| \\ &= |\|v\|-1|\end{aligned}$$

and

$$\begin{aligned}
\|v-u\| &= \langle v-u, v-u\rangle \\
&= \|v\|^2 - \langle v,u\rangle - \langle u,v\rangle + \|u\|^2 \\
&= \|v\|^2 - 2\operatorname{Re}\langle u,v\rangle + 1 \\
&\geq \|v\|^2 - 2|\langle u,v\rangle| + 1 && (*) \\
&\geq \|v\|^2 - 2\|u\|\|v\| + 1 && \text{[by Cauchy–Schwarz inequality]} \quad (**) \\
&= \|v\|^2 - 2\|v\| + 1 = (\|v\|-1)^2 .
\end{aligned}$$

Thus

$$\|v-u\| \geq |\|v\|-1| = \left\| v - \frac{v}{\|v\|} \right\|.$$

Equality holds if and only if equality in both $(*)$ and $(**)$ hold simultaneously. Equality in $(*)$ holds if and only if

$$\begin{aligned}
&\operatorname{Re}\langle u,v\rangle = |\langle u,v\rangle| = \sqrt{\operatorname{Re}\langle u,v\rangle^2 + \operatorname{Im}\langle u,v\rangle^2} \\
\iff &\langle u,v\rangle \in \mathbb{R}_{\geq 0} \\
\iff &k\|v\|^2 \in \mathbb{R}_{\geq 0} \\
\iff &k \in \mathbb{R}_{\geq 0}
\end{aligned}$$

and equality in $(**)$ holds if and only if

$$\begin{aligned}
&u = kv \\
\iff &|k| = \frac{\|kv\|}{\|v\|} = \frac{\|u\|}{\|v\|} = \frac{1}{\|v\|}
\end{aligned}$$

Hence $k = \frac{1}{\|v\|}$, so $u = \frac{v}{\|v\|}$. □

Exercise 7.8 ([Axl24] 6A Q26, polarisation identity)**.** Suppose V is a real inner product space. Prove that

$$\langle u,v\rangle = \frac{\|u+v\|^2 - \|u-v\|^2}{4}.$$

for all $u, v \in V$.

Solution. We have

$$\begin{aligned}
\|u+v\|^2 - \|u-v\|^2 &= \langle u+v, u+v\rangle - \langle u-v, u-v\rangle \\
&= \left(\|u\|^2 + 2\langle u,v\rangle + \|v\|^2\right) - \left(\|u\|^2 - 2\langle u,v\rangle + \|v\|^2\right) \\
&= 4\langle u,v\rangle .
\end{aligned}$$

□

Exercise 7.9 ([Axl24] 6A Q27, polarisation identity)**.** Suppose V is a complex inner product space. Prove that

$$\langle u, v\rangle = \frac{\|u+v\|^2 - \|u-v\|^2 + \|u+iv\|^2 i - \|u-iv\|^2 i}{4}$$

for all $u, v \in V$.

Solution. Similar to previous exercise. □

Exercise 7.10 ([Axl24] 6B Q1)**.** Suppose $\{e_1, \dots, e_m\}$ is a set of vectors in V such that

$$\|a_1 e_1 + \cdots + a_m e_m\|^2 = |a_1|^2 + \cdots + |a_m|^2$$

for all $a_1, \dots, a_m \in \mathbf{F}$. Show that $\{e_1, \dots, e_m\}$ is an orthonormal set of vectors.

Proof. We have

$$\begin{aligned}\left\|\sum_{i=1}^m a_i e_i\right\|^2 &= \left\langle \sum_{i=1}^m a_i e_i, \sum_{i=1}^m a_i e_i \right\rangle \\ &= \sum_{i=1}^m \sum_{j=1}^m a_i \overline{a_j} \left\langle e_i, e_j \right\rangle \\ &= \sum_{i=1}^m |a_i|^2.\end{aligned}$$

For this holds for arbitrary choices of $a_1, \dots, a_m \in \mathbf{F}$, we need to have that

$$\left\langle e_i, e_j \right\rangle = \delta_{ij},$$

which shows that the vectors are orthogonal to each other. To see that each of them has norm 1, we can set $a_k = 1$ and $a_j = 0$ for all $j \neq k$, which gives that $\|e_k\|^2 = |a_k| = 1$, and thus each of the vectors is normalised, completing the proof. □

Exercise 7.11 ([Axl24] 6B Q3)**.** Suppose $\{e_1, \dots, e_m\}$ is an orthonormal set in V and $v \in V$. Prove that

$$\|v\|^2 = |\langle v, e_1\rangle|^2 + \cdots + |\langle v, e_m\rangle|^2 \iff v \in \operatorname{span}(e_1, \dots, e_m).$$

Solution. $\boxed{\Longrightarrow}$ We can decompose v into two parts, one is

$$v_{\text{proj}} = \sum_{i=1}^m \langle v, e_i\rangle e_i,$$

which is the orthogonal projection of v onto the subspace spanned by $e_1, \ldots, e_m$. We claim that $v - v_{\text{proj}}$ is orthogonal to v_{proj}. This can be seen as

$$\begin{aligned}\left\langle v_{\text{proj}}, v - v_{\text{proj}} \right\rangle &= \left\langle \sum_{i=1}^{m} \langle v, e_i \rangle e_i, v - \sum_{i=1}^{m} \langle v, e_i \rangle e_i \right\rangle \\ &= \sum_{i=1}^{m} |\langle v, e_i \rangle|^2 - \sum_{i=1}^{m} |\langle v, e_i \rangle|^2 = 0.\end{aligned}$$

Then by Pythagoras' theorem we have

$$\|v\|^2 = \left\|v_{\text{proj}}\right\|^2 + \left\|v - v_{\text{proj}}\right\|^2$$

where $\|v\|^2 = \left\|v_{\text{proj}}\right\|^2$ and thus $v = v_{\text{proj}}$. Equivalently, $v \in \text{span}(e_1, \ldots, e_m)$.

$\boxed{\Longleftarrow}$ This means that $v = \sum_{i=1}^{m} a_i e_i$. However, we know that $a_i = \langle v, e_i \rangle$, so $\|v\|^2 = \sum_{i=1}^{m} |\langle v, e_i \rangle|^2$ by repeatedly applying Pythagoras' theorem. □

Exercise 7.12 ([Axl24] 6B Q6)**.** Suppose $\{e_1, \ldots, e_n\}$ is an orthonormal basis of V.

(i) Prove that if $v_1, \ldots, v_n$ are vectors in V such that

$$\|e_i - v_i\| < \frac{1}{\sqrt{n}}$$

for each i, then $\{v_1, \ldots, v_n\}$ is a basis of V.

(ii) Show that there exist $v_1, \ldots, v_n \in V$ such that

$$\|e_i - v_i\| \leq \frac{1}{\sqrt{n}}$$

for each i, but $\{v_1, \ldots, v_n\}$ is not linearly independent.

Exercise 7.13 ([Axl24] 6B Q9)**.** Suppose $e_1, \ldots, e_m$ is the result of applying the Gram–Schmidt procedure to a linearly independent set $v_1, \ldots, v_m$ in V. Prove that $\langle v_i, e_i \rangle > 0$ for each $i = 1, \ldots, m$.

Exercise 7.14 ([Axl24] 6B Q10)**.** Suppose $\{v_1, \ldots, v_m\}$ is a linearly independent set in V. Explain why the orthonormal set produced by the formulae of the Gram–Schmidt procedure is the only orthonormal set $\{e_1, \ldots, e_m\}$ in V such that $\langle v_i, e_i \rangle > 0$ and $\text{span}(v_1, \ldots, v_i) = \text{span}(e_1, \ldots, e_i)$ for each $i = 1, \ldots, m$.

Exercise 7.15 ([Axl24] 6B Q13)**.** Show that a set $v_1, \ldots, v_m$ of vectors in V is linearly dependent if and only if the Gram–Schmidt formula produces $u_i = \mathbf{0}$ for some

$i \in \{1, \dots, m\}$.

Exercise 7.16 ([Axl24] 6B Q14)**.** Suppose V is a real inner product space and $v_1, \dots, v_m$ is a linearly independent set of vectors in V. Prove that there exist exactly 2^m orthonormal sets $\{e_1, \dots, e_m\}$ of vectors in V such that

$$\operatorname{span}(v_1, \dots, v_i) = \operatorname{span}(e_1, \dots, e_i)$$

for all $i = 1, \dots, m$.

Exercise 7.17 ([Axl24] 6B Q15)**.** Suppose $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$ are inner products on V such that $\langle u, v \rangle_1 = 0$ if and only if $\langle u, v \rangle_2 = 0$. Prove that there exists $c > 0$ such that $\langle u, v \rangle_1 = c \langle u, v \rangle_2$ for every $u, v \in V$.

This exercise shows that if two inner products have the same pairs of orthogonal vectors, then each of the inner products is a scalar multiple of the other inner product.

Exercise 7.18 ([Axl24] 6B Q16)**.** Suppose V is finite-dimensional. Suppose $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$ are inner products on V with corresponding norms $\|\cdot\|_1$ and $\|\cdot\|_2$. Prove that there exists $c > 0$ such that $\|v\|_1 \leq c\|v\|_2$ for every $v \in V$.

Exercise 7.19 ([Axl24] 6B Q17)**.** Suppose V is a complex finite-dimensional vector space. Prove that if T is an operator on V such that 1 is the only eigenvalue of T and $\|Tv\| \leq \|v\|$ for all $v \in V$, then T is the identity operator.

6C 1 4 5 6 7 8 9 10 11 12 14

Exercise 7.20 ([Axl24] 6C Q1)**.**

Exercise 7.21 ([Axl24] 6C Q4)**.**

Exercise 7.22 ([Axl24] 6C Q5)**.**

Exercise 7.23 ([Axl24] 6C Q6)**.**

Exercise 7.24 ([Axl24] 6C Q7)**.**

Exercise 7.25 ([Axl24] 6C Q8)**.** Suppose $U \leq V$ is finite-dimensional, $v \in V$. Define a linear functional $\phi : U \to \mathbf{F}$ by

$$\phi(u) = \langle u, v \rangle$$

for all $u \in U$. By the Riesz representation theorem applied to the inner product space U, there exists a unique vector $w \in U$ such that

$$\phi(u) = \langle u, w \rangle$$

for all $u \in U$. Show that $w = P_U v$.

Solution. For each $u \in U$, we have $\langle u, v \rangle = \langle u, w \rangle$. Write

$$\langle u, v \rangle = \langle u, P_U v + (v - P_U v) \rangle .$$

Since $v - P_U v \in U^\perp$, this implies $\langle u, v - P_U v \rangle = 0$. Thus

$$\langle u, v \rangle = \langle u, P_U v + (v - P_U v) \rangle = \langle u, P_U v \rangle .$$

By the uniqueness of w, we must have $w = P_U v$. □

Exercise 7.26 ([Axl24] 6C Q9)**.**

Exercise 7.27 ([Axl24] 6C Q10)**.** Suppose V is finite-dimensional, and $P \in \mathscr{L}(V)$ is such that $P^2 = P$ and

$$\|Pv\| \leq \|v\| \quad (v \in V).$$

Prove that there exists a subspace U of V such that $P = P_U$.

Solution. □

Exercise 7.28 ([Axl24] 6C Q11)**.**

Exercise 7.29 ([Axl24] 6C Q12)**.**

Exercise 7.30 ([Axl24] 6C Q14)**.**

8 Operators on Inner Product Spaces

8.1 Self-Adjoint and Normal Operators

Adjoints

Definition 8.1 (Adjoint). The ***adjoint*** of $T \in \mathscr{L}(V,W)$ is the function $T^* : W \to V$ such that

$$\langle Tv, w\rangle = \langle v, T^*w\rangle \quad (v \in V,\, w \in W).$$

Remark. You can remember this as "flipping T to the other side".

We need to check that this definition makes sense. Suppose $T \in \mathscr{L}(V,W)$, fix $w \in W$. Consider the linear functional on V which maps

$$v \mapsto \langle Tv, w\rangle .$$

By the Riesz representation theorem, there exists a unique vector in V such that this linear functional is given by taking the inner product with it; we call this unique vector T^*w.

Notation. In the equation above, the inner product on the LHS takes place in W, and the inner product on the right takes place in V. However, we use the same notation $\langle \cdot, \cdot \rangle$ for both inner products.

Lemma. *The adjoint of a linear map is a linear map.*

Proof. Suppose $T \in \mathscr{L}(V,W)$. We want to show that $T^* \in \mathscr{L}(W,V)$.

(i) Let $v \in V$, $w_1, w_2 \in W$. Then

$$\begin{aligned}\langle Tv, w_1 + w_2\rangle &= \langle Tv, w_1\rangle + \langle Tv, w_2\rangle \\ &= \langle v, T^*w_1\rangle + \langle v, T^*w_2\rangle \\ &= \langle v, T^*w_1 + T^*w_2\rangle .\end{aligned}$$

By the Riesz representation theorem, $T^*(w_1 + w_2) = T^*w_1 + T^*w_2$.

(ii) Let $v \in V$, $\lambda \in \mathbf{F}$, $w \in W$. Then

$$\langle Tv, \lambda w \rangle = \overline{\lambda} \langle Tv, w \rangle = \overline{\lambda} \langle v, T^* w \rangle = \langle v, \lambda T^* w \rangle .$$

By the Riesz representation theorem, $T^*(\lambda w) = \lambda T^* w$.

□

Remark. To compute T^*, start with a formula for $\langle Tv, w \rangle$, then manipulate it to get *only* v in the first slot; the entry in the second slot will then be $T^* w$.

Lemma 8.2 (Properties of adjoint). *Suppose $T \in \mathscr{L}(V,W)$.*

(i) $(S+T)^* = S^* + T^*$ *for all* $S \in \mathscr{L}(V,W)$.

(ii) $(\lambda T)^* = \overline{\lambda} T^*$ *for all* $\lambda \in \mathbf{F}$.

(iii) $(T^*)^* = T$.

(iv) $(ST)^* = T^* S^*$ *for all* $S \in \mathscr{L}(W,U)$, *where* U *is a finite-dimensional inner product space over* $\mathbf{F}$.

(v) $I^* = I$, *where* I *is the identity operator on* V.

(vi) If T *is invertible, then* T^* *is invertible, and* $(T^*)^{-1} = (T^{-1})^*$.

Proof. Let $v \in V$, $w \in V$.

(i) If $S \in \mathscr{L}(V,W)$, then

$$\begin{aligned} \langle (S+T)v, w \rangle &= \langle Sv, w \rangle + \langle Tv, w \rangle \\ &= \langle v, S^* w \rangle + \langle v, T^* w \rangle \\ &= \langle v, (S^* + T^*) w \rangle . \end{aligned}$$

Hence $(S+T)^* w = (S^* + T^*) w$.

(ii) Let $\lambda \in \mathbf{F}$, then

$$\langle (\lambda T)v, w \rangle = \lambda \langle Tv, w \rangle = \lambda \langle v, T^* w \rangle = \left\langle v, \overline{\lambda} T^* w \right\rangle .$$

Hence $(\lambda T)^* w = \overline{\lambda} T^* w$.

(iii) We have

$$\langle T^* w, v \rangle = \overline{\langle v, T^* w \rangle} = \overline{\langle Tv, w \rangle} = \langle w, Tv \rangle .$$

Hence $(T^*)^* v = Tv$.

(iv) Let $S \in \mathscr{L}(W,U)$, $u \in U$. Then

$$\langle (ST)v, u\rangle = \langle S(Tv), u\rangle = \langle Tv, S^*u\rangle = \langle v, T^*(S^*u)\rangle .$$

Hence $(ST)^*u = T^*S^*u$.

(v) Let $u \in V$. Then

$$\langle Iu, v\rangle = \langle u, v\rangle .$$

Hence $I^*v = v$.

(vi) Suppose T is invertible. Then $T^{-1}T = I$. Taking adjoints of both sides and applying (iv) and (v) gives

$$T^*(T^{-1})^* = I.$$

Similarly, the equation $TT^{-1} = I$ implies

$$(T^{-1})^*T^* = I.$$

Hence $(T^{-1})^*$ is the inverse of T^*.

□

If $\mathbf{F} = \mathbb{R}$, then the map $T \mapsto T^*$ is a linear map from $\mathscr{L}(V,W)$ to $\mathscr{L}(W,V)$, as follows from (i) and (ii). However if $\mathbf{F} = \mathbb{C}$, then this map is not linear due to the complex conjugate in (ii).

The next result shows the relationship between the kernel and image of a linear map and its adjoint.

Lemma 8.3 (Kernel and image of T^*). *Suppose $T \in \mathscr{L}(V,W)$. Then*

(i) $\ker T^* = (\operatorname{im} T)^\perp$

(ii) $\operatorname{im} T^* = (\ker T)^\perp$

(iii) $\ker T = (\operatorname{im} T^*)^\perp$

(iv) $\operatorname{im} T = (\ker T^*)^\perp$

Proof.

(i) Let $w \in W$. Then

$$\begin{aligned} w \in \ker T^* &\iff T^*w = 0 \\ &\iff \langle v, T^*w\rangle = 0 \quad \forall v \in V \\ &\iff \langle Tv, w\rangle = 0 \quad \forall v \in V \\ &\iff w \in (\operatorname{im} T)^\perp. \end{aligned}$$

Hence $\ker T^* = (\operatorname{im} T)^\perp$.

(ii) Replace T with T^* in (iv).

(iii) Replace T with T^* in (i), and use the fact that $(T^*)^* = T$.

(iv) Take the orthogonal complement of both sides of (i), and use the fact that $U = (U^\perp)^\perp$ if $U \leq V$.

□

Property (iv) of the above result has a useful consequence; sometimes it is easier to determine that a linear map is injective, rather than surjective.

Corollary 8.4 (Condition for surjectivity)**.** *Suppose $T \in \mathscr{L}(V,W)$. Then*

$$T \text{ is surjective} \iff T^* \text{ is injective.}$$

Proof. We have

$$\begin{aligned} T \text{ is surjective} &\iff \operatorname{im} T = W \\ &\iff (\ker T^*)^\perp = W \\ &\iff \ker T^* = \{\mathbf{0}\} \\ &\iff T^* \text{ is injective.} \end{aligned}$$

□

As we will soon see, the next definition is closely related to the matrix of the adjoint of a linear map.

Definition 8.5 (Conjugate transpose)**.** The ***conjugate transpose*** of a $m \times n$ matrix A is the $n \times m$ matrix A^* obtained by taking the complex conjugate of each entry of $A^\top$.

That is, $(A^*)_{ij} := \overline{A_{ji}}$.

The next result shows how to compute the matrix of T^* from the matrix of T.

Lemma 8.6. *Let $T \in \mathscr{L}(V,W)$. Suppose $\{e_1,\ldots,e_n\}$ is an orthonormal basis of V, and $\{f_1,\ldots,f_m\}$ is an orthonormal basis of W. Then*

$$\mathscr{M}(T^*) = \mathscr{M}(T)^*.$$

Remark. $\mathscr{M}(T;\{e_1,\ldots,e_n\},\{f_1,\ldots,f_m\})$ and $\mathscr{M}(T^*;\{f_1,\ldots,f_m\},\{e_1,\ldots,e_n\})$.

Proof. Denote $A = \mathscr{M}(T)$, $B = \mathscr{M}(T^*)$. To show that $B = A^*$, we need to show that $B_{ij} = (A^*)_{ij} = \overline{A_{ji}}$.

Since $\{f_1, \dots, f_m\}$ is an orthonormal basis of W, we can write

$$Te_j = \langle Te_j, f_1 \rangle f_1 + \cdots + \langle Te_j, f_m \rangle f_m$$

where $j = 1, \dots, n$. Thus $A_{ij} = \langle Te_j, f_i \rangle$.

Replacing T with T^*, and interchanging $\{e_1, \dots, e_n\}$ and $\{f_1, \dots, f_m\}$ gives

$$B_{ij} = \langle T^* f_j, e_i \rangle = \langle f_j, Te_i \rangle = \overline{\langle Te_i, f_j \rangle} = \overline{A_{ji}}.$$

Hence $\mathscr{M}(T^*) = \mathscr{M}(T)^*$. □

Self-Adjoint Operators

Definition 8.7 (Self-adjoint operator)**.** An operator is ***self-adjoint*** if it equals its adjoint.

That is, $T \in \mathscr{L}(V)$ is self-adjoint if $T = T^*$, i.e.,

$$\langle Tv, w \rangle = \langle v, Tw \rangle \quad (v, w \in V).$$

Lemma 8.8. *Every eigenvalue of a self-adjoint operator is real.*

Idea. To show a number is real, we need to show that it equals its conjugate.

Proof. Suppose $T \in \mathscr{L}(V)$ is self-adjoint. Let λ be an eigenvalue of T, with corresponding eigenvector $v \in V \setminus \{\mathbf{0}\}$. Then $Tv = \lambda v$. We have

$$\lambda \|v\|^2 = \langle \lambda v, v \rangle = \langle Tv, v \rangle = \langle v, Tv \rangle = \langle v, \lambda v \rangle = \overline{\lambda} \|v\|^2.$$

Since $v \neq \mathbf{0}$, we have $\lambda = \overline{\lambda}$, which means that λ is real. □

Lemma 8.9. *Suppose V is a* complex *inner product space, and $T \in \mathscr{L}(V)$. Then*

$$\langle Tv, v \rangle = 0 \quad \forall v \in V \iff T = 0.$$

Remark. This result does not hold for real inner product spaces. For instance, the operator $T \in \mathscr{L}(\mathbb{R}^2)$ that is a counterclockwise rotation of 90° around the origin; thus $T(x, y) = (-y, x)$. Notice that Tv is orthogonal to v for every $v \in \mathbb{R}^2$, even though $T \neq 0$.

Proof.

$\boxed{\Longleftarrow}$ Suppose $T = 0$. If $u, w \in V$, then

$$\langle Tu, w\rangle = \frac{\langle T(u+w), u+w\rangle - \langle T(u-w), u-w\rangle}{4} + \frac{\langle T(u+iw), u+iw\rangle - \langle T(u-iw), u-iw\rangle}{4} i.$$

Note that each term on the RHS is of the form $\langle Tv, v\rangle$ for appropriate $v \in V$.

$\boxed{\Longrightarrow}$ Suppose $\langle Tv, v\rangle = 0$ for every $v \in V$.

Then the equation above implies that $\langle Tu, w\rangle = 0$ for all $u, w \in V$. Taking $w = Tu$ for every $u \in V$, we obtain $Tu = \mathbf{0}$ for every $u \in V$. Hence $T = 0$ as desired. □

The next result provides a characterisation of self-adjoint operators over $\mathbb{C}$.

Lemma 8.10. *Suppose V is a* complex *inner product space, and $T \in \mathscr{L}(V)$. Then*

$$T \text{ is self-adjoint} \iff \langle Tv, v\rangle \in \mathbb{R} \quad \forall v \in V.$$

Remark. This result does not hold for real inner product spaces, by considering any operator on a real inner product space that is not self-adjoint.

Proof. If $v \in V$, then

$$\langle T^* v, v\rangle = \overline{\langle v, T^* v\rangle} = \overline{\langle Tv, v\rangle}. \tag{I}$$

where the second equality follows since T is self-adjoint. Thus

$$\begin{aligned}
T \text{ is self-adjoint} &\iff T = T^* \\
&\iff T - T^* = 0 \\
&\iff \langle (T - T^*)v, v\rangle = 0 \quad \forall v \in V && \text{[by 8.9]} \\
&\iff \langle Tv, v\rangle = \langle T^* v, v\rangle \quad \forall v \in V \\
&\iff \langle Tv, v\rangle = \overline{\langle Tv, v\rangle} \quad \forall v \in V && \text{[by (I)]} \\
&\iff \langle Tv, v\rangle \in \mathbb{R} \quad \forall v \in V.
\end{aligned}$$

□

On a real inner product space V, a non-zero operator T might satisfy $\langle Tv, v\rangle = 0$ for all $v \in V$. However, the next result shows that this cannot happen for a self-adjoint operator.

Lemma 8.11. *Suppose T is a self-adjoint operator on V. Then*

$$\langle Tv, v\rangle = 0 \quad \forall v \in V \iff T = 0.$$

Proof. We have already proved this (without the hypothesis that T is self-adjoint) when V is a complex inner product space (see 8.9). Thus we can assume that V is a real inner product space.

$\boxed{\Longleftarrow}$ Let $u, v \in V$, then

$$\langle Tu, w\rangle = \frac{\langle T(u+w), u+w\rangle - \langle T(u-w), u-w\rangle}{4} \tag{I}$$

as can be proved by computing the RHS using $\langle Tw, u\rangle = \langle w, Tu\rangle = \langle Tu, w\rangle$, where the first equality holds because T is self-adjoint, and the second equality holds because we are working in a real inner product space.

$\boxed{\Longrightarrow}$ Suppose $\langle Tv, v\rangle = 0$ for every $v \in V$.

Since each term on the RHS of (I) is of the form $\langle Tv, v\rangle$ for appropriate v, this implies that $\langle Tu, w\rangle = 0$ for all $u, w \in V$. Thus taking $w = Tu$, we obtain $Tu = \mathbf{0}$ for every $u \in V$. Hence $T = 0$, as desired. □

Normal Operators

Definition 8.12 (Normal operator)**.** An operator is ***normal*** if it commutes with its adjoint.

That is, $T \in \mathscr{L}(V)$ is normal if $TT^* = T^*T$.

Remark. Every self-adjoint operator is normal, but not vice versa.

The next result provides a useful characterisation of normal operators.

Lemma 8.13 (Characterisation of normal operators)**.** *Suppose $T \in \mathscr{L}(V)$. Then*

$$T \text{ is normal} \iff \|Tv\| = \|T^*v\| \quad \forall v \in V.$$

Proof. We have

$$\begin{aligned}
T \text{ is normal} &\iff TT^* = T^*T \\
&\iff T^*T - TT^* = 0 \\
&\iff \langle (T^*T - TT^*)v, v\rangle = 0 \quad \forall v \in V \\
&\iff \langle T^*Tv, v\rangle = \langle TT^*v = v\rangle \quad \forall v \in V \\
&\iff \langle Tv, Tv\rangle = \langle T^*v, T^*v\rangle \quad \forall v \in V \\
&\iff \|Tv\| = \|T^*v\| \quad \forall v \in V.
\end{aligned}$$

□

The next result presents several consequences of the result above.

Lemma 8.14. *Suppose $T \in \mathscr{L}(V)$ is normal.*

(i) $\ker T = \ker T^*$.

(ii) $\operatorname{im} T = \operatorname{im} T^*$.

(iii) $V = \ker T \oplus \operatorname{im} T$.

(iv) $T - \lambda I$ *is normal for every* $\lambda \in \mathbf{F}$.

(v) If $v \in V$ *and* $\lambda \in \mathbf{F}$, *then* $Tv = \lambda v \iff T^*v = \overline{\lambda} v$.

Proof.

(i) Let $v \in V$. Then

$$\begin{aligned} v \in \ker T &\iff Tv = \mathbf{0} \\ &\iff \|Tv\| = 0 \\ &\iff \|T^*v\| = 0 && \text{[by 8.13]} \\ &\iff T^*v = \mathbf{0} \\ &\iff v \in \ker T^* \end{aligned}$$

(ii) We have

$$\begin{aligned} \operatorname{im} T &= (\ker T^*)^\perp && \text{[by 8.3]} \\ &= (\ker T)^\perp && \text{[by (i)]} \\ &= \operatorname{im} T^* && \text{[by 8.3]} \end{aligned}$$

(iii) We have

$$\begin{aligned} V &= (\ker T) \oplus (\ker T)^\perp \\ &= \ker T \oplus \operatorname{im} T^* && \text{[by 8.3]} \\ &= \ker T \oplus \operatorname{im} T && \text{[by (ii)]} \end{aligned}$$

(iv) Let $\lambda \in \mathbf{F}$, then

$$\begin{aligned} (T - \lambda I)(T - \lambda I)^* &= (T - \lambda I)(T^* - \overline{\lambda} I) \\ &= TT^* - \overline{\lambda} T - \lambda T^* + |\lambda|^2 I \\ &= T^*T - \overline{\lambda} T - \lambda T^* + |\lambda|^2 I \\ &= (T^* - \overline{\lambda} I)(T - \lambda I) \\ &= (T - \lambda I)^*(T - \lambda I). \end{aligned}$$

Thus $T - \lambda I$ commutes with its adjoint. Hence $T - \lambda I$ is normal.

(v) Let $v \in V$, $\lambda \in \mathbf{F}$. Then $Tv = \lambda v$ if and only if

$$\begin{aligned} 0 &= \|(T - \lambda I)v\| \\ &= \|(T - \lambda I)^* v\| && \text{[by (iv) and 8.13]} \\ &= \left\|(T^* - \overline{\lambda} I)v\right\| \end{aligned}$$

if and only if $T^*v = \overline{\lambda} v$.

□

Property (v) implies that if T is normal, then T and T^* have the same eigenvectors.

Proposition 8.15. *Suppose $T \in \mathscr{L}(V)$ is normal. Then the eigenvectors of T corresponding to distinct eigenvalues are orthogonal.*

Proof. Let α, β be distinct eigenvalues of T, with corresponding eigenvectors u, v. Thus $Tu = \lambda u$ and $Tv = \beta v$. By 8.14, $Tv = \beta v \iff T^*v = \overline{\beta} v$. Thus

$$\begin{aligned} (\alpha - \beta)\langle u, v\rangle &= \langle \alpha u, v\rangle - \langle \beta u, v\rangle \\ &= \langle \alpha u, v\rangle - \left\langle u, \overline{\beta} v\right\rangle \\ &= \langle Tu, v\rangle - \langle u, T^*v\rangle \\ &= 0. \end{aligned}$$

Since $\alpha \neq \beta$, the equation above implies that $\langle u, v\rangle = 0$. Hence u and v are orthogonal, as desired. □

Proposition 8.16. *Suppose $\mathbf{F} = \mathbb{C}$ and $T \in \mathscr{L}(V)$. Then T is normal if and only if there exist commuting self-adjoint operators A and B such that $T = A + iB$.*

Proof.

$\boxed{\Longrightarrow}$ Suppose T is normal.

Claim. $A = \dfrac{T + T^*}{2}$, $B = \dfrac{T - T^*}{2}$.

Then A and B are self-adjoint, and $T = A + iB$. A quick computation shows that

$$AB - BA = \frac{T^*T - TT^*}{2i}.$$

Since T is normal, the RHS equals 0. Thus $AB = BA$, so A and B commute.

$\boxed{\Longleftarrow}$ Suppose there exist commuting self-adjoint operators A and B such that

$$T = A + iB.$$

Then the adjoint is

$$T^* = A - iB.$$

Solving for A and B gives

$$A = \frac{T + T^*}{2}, \quad B = \frac{T - T^*}{2}.$$

This implies that

$$AB - BA = \frac{T^*T - TT^*}{2i}.$$

Since A and B commute, the LHS equals 0, so T is normal, as desired. □

8.2 Spectral Theorem

Recall that 6.30 states that an operator on V has a diagonal matrix with respect to a basis if and only if the basis consists of eigenvectors of the operator. Now we are interested in the question of which operators on V is there an *orthonormal* basis of V with respect to which the operator has a diagonal matrix.

By 6.30, this question is equivalent to the question of which operators on V is there an orthonormal basis of V consisting of eigenvectors of T.

The *spectral theorem* will answer these questions, depending on whether $\mathbf{F} = \mathbb{R}$ or $\mathbf{F} = \mathbb{C}$.

Real Spectral Theorem

To prove the real spectral theorem, we will need two preliminary results. These preliminary results hold on both real and complex inner product spaces, but they are not needed for the proof of the complex spectral theorem.

Lemma 8.17 (Invertible quadratic expressions). *Suppose $T \in \mathscr{L}(V)$ is self-adjoint and $b, c \in \mathbb{R}$ are such that $b^2 < 4c$. Then*

$$T^2 + bT + cI$$

is an invertible operator.

Proof. It suffices to prove that $T^2 + bT + cI$ is injective.

Let $v \in V$ be a non-zero vector. Then

$$\begin{aligned}
&\big\langle (T^2 + bT + cI)v, v \big\rangle \\
&= \big\langle T^2 v, v \big\rangle + b \langle Tv, v \rangle + c \langle v, v \rangle \\
&= \langle Tv, Tv \rangle + b \langle Tv, v \rangle + c\|v\|^2 \\
&\geq \|Tv\|^2 - |b|\|Tv\|\|v\| + c\|v\|^2 && \text{[by Cauchy–Schwarz inequality]} \\
&= \left(\|Tv\| - \frac{|b|\|v\|}{2} \right)^2 + \left(c - \frac{b^2}{4} \right) \|v\|^2 > 0 && \text{[completing the square]}
\end{aligned}$$

This implies that $(T^2 + bT + cI)v \neq \mathbf{0}$ for all $v \neq \mathbf{0}$. Thus $\ker(T^2 + bT + cI) = \{\mathbf{0}\}$, so $T^2 + bT + cI$ is injective. □

Lemma 8.18 (Minimal polynomial of self-adjoint operator). *Suppose $T \in \mathscr{L}(V)$ is*

self-adjoint. Then the minimal polynomial of T equals

$$(z-\lambda_1)\cdots(z-\lambda_m)$$

for some $\lambda_1,\ldots,\lambda_m \in \mathbb{R}$.

Proof. Since we are dealing with polynomials, we deal with the cases where $\mathbf{F}=\mathbb{C}$ and $\mathbf{F}=\mathbb{R}$ separately.

$\mathbf{F}=\mathbb{C}$: By 6.14, the zeros of the minimal polynomial of T are the eigenvalues of T. By 8.8, all eigenvalues of T are real. Thus the second version of the fundamental theorem of algebra (5.11) tells us that the minimal polynomial of T has the desired form.

$\mathbf{F}=\mathbb{R}$: By the factorisation of a polynomial over $\mathbb{R}$ (5.14), the minimal polynomial of T equals

$$p(z)=(z-\lambda_1)\cdots(z-\lambda_m)(z^2+b_1z+c_1)\cdots(z^2+b_Nz+c_N). \tag{I}$$

for some $\lambda_1,\ldots,\lambda_m \in \mathbb{R}$ and $b_1,\ldots,b_N,c_1,\ldots,c_N \in \mathbb{R}$ with ${b_i}^2 < 4c_i$ for each $i=1,\ldots,N$.

Claim. $N=0$. (This means there are no terms of the form $z^2+b_iz+c_i$.)

Since p is the minimal polynomial,

$$p(T)=(T-\lambda_1 I)\cdots(T-\lambda_m I)(T^2+b_1T+c_1I)\cdots(T^2+b_NT+c_NI)=0.$$

Suppose, for a contradiction, that $N>0$.

By 8.17, $T^2+b_NT+c_NI$ is invertible, so we multiply both sides of the equation above on the right by $(T^2+b_NT+c_NI)^{-1}$. We then obtain a polynomial expression of T that equals 0. However this polynomial would have degree two less than the degree of p in (I), contradicting the minimality of $\deg p$.

Hence $N=0$, which means that the minimal polynomial in (I) has the form $(z-\lambda_1)\cdots(z-\lambda_m)$, as desired.

□

The next result, which gives a complete description of the self-adjoint operators on a real inner product space, is one of the major theorems in linear algebra.

Theorem 8.19 (Real spectral theorem). *Suppose $\mathbf{F}=\mathbb{R}$ and $T \in \mathscr{L}(V)$. Then the following are equivalent:*

(i) T is self-adjoint.

(ii) T has a diagonal matrix with respect to some orthonormal basis of V.

(iii) V has an orthonormal basis consisting of eigenvectors of T.

Proof.

(i) $\Longrightarrow$ (ii) Suppose T is self-adjoint. Our results on minimal polynomials, specifically 7.21 and 8.18, imply that T has an upper-triangular matrix $\mathscr{M}(T)$ with respect to some orthonormal basis of V. With respect to this orthonormal basis, $\mathscr{M}(T^*) = \mathscr{M}(T)^\top$.

However, $T^* = T$. Thus $\mathscr{M}(T)^\top = \mathscr{M}(T)$. Since $\mathscr{M}(T)$ is upper-triangular, this means that all entries of the matrix above and below the diagonal are 0. Hence the matrix of T is a diagonal matrix with respect to the orthonormal basis.

(ii) $\Longrightarrow$ (i) Suppose T has a diagonal matrix $\mathscr{M}(T)$ with respect to some orthonormal basis of V.

That diagonal matrix equals its transpose. Thus with respect to that basis, $\mathscr{M}(T^*) = \mathscr{M}(T)$. Hence $T^* = T$, so T is self-adjoint.

(ii) $\Longleftrightarrow$ (iii) This follows from the definitions. □

Complex Spectral Theorem

The next result gives a complete description of the normal operators on a complex inner product space.

Theorem 8.20 (Complex spectral theorem). *Suppose $\mathbf{F} = \mathbb{C}$ and $T \in \mathscr{L}(V)$. Then the following are equivalent:*

(i) T is normal.

(ii) T has a diagonal matrix with respect to some orthonormal basis of V.

(iii) V has an orthonormal basis consisting of eigenvectors of T.

Proof.

(i) $\Longrightarrow$ (ii) Suppose T is normal.

By Schur's theorem, there exists an orthonormal basis $\{e_1, \ldots, e_n\}$ of V, with respect to which T has an upper-triangular matrix:

$$\mathscr{M}(T; \{e_1, \ldots, e_n\}) = \begin{pmatrix} A_{11} & \cdots & A_{1n} \\ & \ddots & \vdots \\ 0 & & A_{nn} \end{pmatrix}.$$

Claim. $\mathcal{M}(T)$ is a diagonal matrix.

We see that

$$\|Te_1\|^2 = |A_{11}|^2$$
$$\|T^*e_1\|^2 = |A_{11}|^2 + |A_{12}|^2 + \cdots + |A_{1n}|^2.$$

Since T is normal, $\|Te_1\| = \|T^*e_1\|$ (by 8.13). Thus the two equations above imply that all entries in the first row of $\mathcal{M}(T)$, except possibly the first entry A_{11}, equal 0.

Now since $A_{12} = 0$, we have

$$\|Te_2\|^2 = |A_{22}|^2$$
$$\|T^*e_2\|^2 = |A_{22}|^2 + |A_{23}|^2 + \cdots + |A_{2n}|^2.$$

Since T is normal, $\|Te_2\| = \|T^*e_2\|$. Thus the two equations above imply that all entries in the second row of $\mathcal{M}(T)$, except possibly the diagonal entry A_{22}, equal 0.

Continuing in this fashion, we see that all non-diagonal entries in $\mathcal{M}(T)$ equal 0. Thus $\mathcal{M}(T)$ is a diagonal matrix.

(ii) $\Longrightarrow$ (i) Suppose T has a diagonal matrix with respect to some orthonormal basis of V. The matrix of T^* (with respect to the same basis) is obtained by taking the conjugate transpose of the matrix of T; hence T^* also has a diagonal matrix. Any two diagonal matrices commute; thus T commutes with T^*, which means that T is normal.

(ii) $\Longleftrightarrow$ (iii) This follows from 6.30. □

8.3 Positive Operators

Definition 8.21 (Positive operator)**.** An operator $T \in \mathscr{L}(V)$ is ***positive*** if

(i) T is self-adjoint, and

(ii) $\langle Tv, v\rangle \geq 0$ for all $v \in V$.

Remark. If V is a complex vector space, then requirement that T be self-adjoint can be dropped from the definition above (by 8.10).

Example.

- If $U \leq V$, then the orthogonal projection P_U is a positive operator.
- If $T \in \mathscr{L}(V)$ is self-adjoint and $b, c \in \mathbb{R}$ are such that $b^2 \leq 4c$, then $T^2 + bT + cI$ is a positive operator.
- If $\mathscr{M}(T)$ is a diagonal matrix with non-negative entries on the diagonal, then T is a positive operator.

Definition 8.22 (Squared root)**.** An operator R is a ***square root*** of an operator T if $R^2 = T$.

Lemma 8.23 (Characterisation of positive operators)**.** *Let $T \in \mathscr{L}(V)$. Then the following are equivalent:*

(i) T is a positive operator.

(ii) T is self-adjoint and all eigenvalues of T are non-negative.

(iii) With respect to some orthonormal basis of V, the matrix of T is a diagonal matrix with only non-negative numbers on the diagonal.

(iv) T has a positive square root.

(v) T has a self-adjoint square root.

(vi) $T = R^*R$ for some $R \in \mathscr{L}(V)$.

Proof.

$\boxed{\text{(i)} \implies \text{(ii)}}$ Suppose T is positive. Then T is self-adjoint.

Let λ is an eigenvalue of T, with corresponding eigenvector v. Then

$$0 \leq \langle Tv, v\rangle = \langle \lambda v, v\rangle = \lambda \langle v, v\rangle .$$

Since $\langle v, v\rangle \geq 0$, we must have $\lambda \geq 0$.

(ii) $\Longrightarrow$ (iii) Suppose T is self-adjoint, and all eigenvalues of T are non-negative.

Since T is self-adjoint, by the spectral theorem (8.19 and 8.20), there exists an orthonormal basis $\{e_1, \ldots, e_n\}$ of V consisting of eigenvectors of T.

Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of T corresponding to $e_1, \ldots, e_n$; thus each $\lambda_i \geq 0$.

The matrix of T with respect to $\{e_1, \ldots, e_n\}$ is the diagonal matrix with $\lambda_1, \ldots, \lambda_n$ on the diagonal.

(iii) $\Longrightarrow$ (iv) Suppose $\{e_1, \ldots, e_n\}$ is an orthonormal basis of V such that the matrix of T with respect to this basis is a diagonal matrix, with non-negative $\lambda_1, \ldots, \lambda_n$ on the diagonal.

By the linear map lemma, there exists $R \in \mathscr{L}(V)$ such that

$$Re_i = \sqrt{\lambda_i} e_i$$

for $i = 1, \ldots, n$.

Claim. R is a positive square root of T.

As you should verify, R is a positive operator. Furthermore,

$$R^2 e_i = \lambda_i e_i = Te_i \quad (i = 1, \ldots, n).$$

Thus $R^2 = T$. Hence R is a positive square root of T.

(iv) $\Longrightarrow$ (v) Every positive operator is self-adjoint (by definition of positive operator).

(v) $\Longrightarrow$ (vi) Suppose T has a self-adjoint square root. Then there exists a self-adjoint operator R on V such that $T = R^2$. Then $T = R^* R$ (since $R^* = R$).

(vi) $\Longrightarrow$ (i) Let $R \in \mathscr{L}(V)$ be such that $T = R^* R$. Then

$$T^* = (R^* R)^* = R^* (R^*)^* = R^* R = T.$$

Hence T is self-adjoint. Now for every $v \in V$,

$$\langle Tv, v\rangle = \langle R^* Rv, v\rangle = \langle Rv, Rv\rangle \geq 0.$$

Thus T is positive. □

Every non-negative number has a unique non-negative square root. The next result shows that positive operators enjoy a similar property.

Lemma 8.24. *Every positive operator on V has a unique positive square root.*

Remark. A positive operator can have infinitely many square roots, although only one of them can be positive. For example, the identity operator on V has infinitely many square roots if $\dim V > 1$.

Proof.

Existence This follows from 8.23.

Uniqueness Suppose $T \in \mathscr{L}(V)$ is positive. Suppose $v \in V$ is an eigenvector of T. Hence there exists a real number $\lambda \geq 0$ such that $Tv = \lambda v$.

Let R be a positive square root of T. We will prove that $Rv = \sqrt{\lambda}v$. This will imply that the behaviour of R on the eigenvectors of T is uniquely determined. Since there is a basis of V consisting of eigenvectors of T (by the spectral theorem), this will imply that R is uniquely determined.

To prove that $Rv = \sqrt{\lambda}v$, note that the spectral theorem asserts that there is an orthonormal basis $\{e_1, \ldots, e_n\}$ of V consisting of eigenvectors of R. Since R is a positive operator, all its eigenvalues are non-negative. Thus there exist non-negative numbers $\lambda_1, \ldots, \lambda_n$ such that $Re_i = \sqrt{\lambda_i}e_i$ for each $i = 1, \ldots, n$.

Since $\{e_1, \ldots, e_n\}$ is a basis of V, we can write

$$v = a_1 e_1 + \cdots + a_n e_n$$

for some $a_1, \ldots, a_n \in \mathbf{F}$. Thus

$$Rv = a_1\sqrt{\lambda_1}e_1 + \cdots + a_n\sqrt{\lambda_n}e_n.$$

Hence

$$\lambda v = Tv = R^2 v = a_1\lambda_1 e_1 + \cdots + a_n\lambda_n e_n.$$

The equation above implies that

$$a_1\lambda e_1 + \cdots + a_n\lambda e_n = a_1\lambda_1 e_1 + \cdots + a_n\lambda_n e_n.$$

Thus $a_i(\lambda - \lambda_i) = 0$ for each $i = 1, \ldots, n$. Hence

$$v = \sum_{\{i|\lambda_i = \lambda\}} a_i e_i.$$

Thus

$$Rv = \sum_{\{i|\lambda_i = \lambda\}} a_i\sqrt{\lambda}e_i = \sqrt{\lambda}v$$

as desired. □

Notation. For a positive operator T, let $\sqrt{T}$ denotes the unique positive square root of T.

Corollary 8.25. *Suppose T is a positive operator on V, and $v \in V$ is such that $\langle Tv, v\rangle = 0$. Then $Tv = \mathbf{0}$.*

Proof. We have

$$0 = \langle Tv, v\rangle = \left\langle \sqrt{T}\sqrt{T}v, v \right\rangle = \left\langle \sqrt{T}v, \sqrt{T}v \right\rangle = \left\| \sqrt{T}v \right\|^2.$$

Hence $\sqrt{T}v = \mathbf{0}$. Thus $Tv = \sqrt{T}\left(\sqrt{T}v\right) = \mathbf{0}$, as desired. □

8.4 Isometries, Unitary Operators, and Matrix Factorisation

Isometries

Linear maps that preserve norms are sufficiently important to deserve a name.

Definition 8.26 (Isometry). We say $S \in \mathscr{L}(V,W)$ is an ***isometry*** if

$$\|Sv\| = \|v\| \quad (v \in V).$$

Lemma. *Every isometry is injective.*

Proof. Let $S \in \mathscr{L}(V,W)$ be an isometry. Then

$$v \in \ker S \iff Sv = \mathbf{0} \iff \|v\| = \|Sv\| = 0 \iff v = \mathbf{0}.$$

Hence $\ker S = \{\mathbf{0}\}$. □

Lemma 8.27 (Characterisation of isometries). *Suppose $S \in \mathscr{L}(V,W)$. Suppose $\{e_1,\dots,e_n\}$ is an orthonormal basis of V, and $\{f_1,\dots,f_m\}$ is an orthonormal basis of W. Then the following are equivalent:*

(i) *S is an isometry.*

(ii) *$S^*S = I$.*

(iii) *$\langle Su, Sv\rangle = \langle u, v\rangle$ for all $u, v \in V$.* *(preserves iner products)*

(iv) *$\{Se_1,\dots,Se_n\}$ is orthonormal in W.* *(maps orthonormal basis to orthonormal set)*

(v) *The columns of $\mathscr{M}(S;\{e_1,\dots,e_n\},\{f_1,\dots,f_m\})$ form an orthonormal set in $\mathbf{F}^m$ with respect to the Euclidean inner product.*

Proof.

(i) $\implies$ (ii) Suppose S is an isometry. Let $v \in V$, then

$$\begin{aligned}\langle (I - S^*S)v, v\rangle &= \langle v, v\rangle - \langle S^*Sv, v\rangle \\ &= \|v\|^2 - \langle Sv, Sv\rangle \\ &= \|v\|^2 - \|Sv\|^2 = 0.\end{aligned}$$

Hence the self-adjoint operator $I - S^*S$ equals 0 (by 7.16). Thus $S^*S = I$.

(ii) $\Longrightarrow$ (iii) Suppose $S^*S = I$. Let $u, v \in V$. Then

$$\langle Su, Sv\rangle = \langle S^*Su, u\rangle = \langle Iu, v\rangle = \langle u, v\rangle .$$

(iii) $\Longrightarrow$ (iv) Suppose $\langle Su, Sv\rangle = \langle u, v\rangle$ for all $u, v \in V$.

If $i, j \in \{1, \ldots, n\}$, then

$$\langle Se_i, Se_j\rangle = \langle e_i, e_j\rangle .$$

Hence $\{Se_1, \ldots, Se_n\}$ is an orthonormal set in W.

(iv) $\Longrightarrow$ (v) Suppose $\{Se_1, \ldots, Se_n\}$ is an orthonormal set in W.

Let $A = \mathscr{M}(S; \{e_1, \ldots, e_n\}, \{f_1, \ldots, f_m\})$. If $j, k \in \{1, \ldots, n\}$, the j-th and k-th columns of A are

$$\begin{pmatrix} A_{1j} \\ \vdots \\ A_{mj} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} A_{1k} \\ \vdots \\ A_{mk} \end{pmatrix}$$

so the Euclidean inner product in $\mathbf{F}^m$ of the two columns is

$$\begin{aligned} \sum_{i=1}^{m} A_{ij}\overline{A_{ik}} &= \left\langle \sum_{i=1}^{m} A_{ij} f_i, \sum_{i=1}^{m} A_{ik} f_i \right\rangle \\ &= \langle Se_j, Se_k\rangle \\ &= \begin{cases} 1 & (j = k) \\ 0 & (j \neq k) \end{cases} \end{aligned} \tag{I}$$

Thus the columns of A form an orthonormal set in $\mathbf{F}^m$.

(v) $\Longrightarrow$ (i) Suppose the columns of the matrix A defined above form an orthonormal set in $\mathbf{F}^m$. Then (I) shows that $\{Se_1, \ldots, Se_n\}$ is an orthonormal set in W.

Let $v \in V$. Then

$$v = \langle v, e_1\rangle e_1 + \cdots + \langle v, e_n\rangle e_n$$

and

$$\|v\|^2 = |\langle v, e_1\rangle|^2 + \cdots + |\langle v, e_n\rangle|^2 .$$

Applying S gives

$$Sv = \langle v, e_1\rangle Se_1 + \cdots + \langle v, e_n\rangle Se_n$$

so

$$\|Sv\|^2 = |\langle v, e_1\rangle|^2 + \cdots + |\langle v, e_n\rangle|^2 .$$

Thus $\|Sv\| = \|v\|$, so S is an isometry. □

Unitary Operators

Definition 8.28 (Unitary operator). We say $S \in \mathscr{L}(V)$ is ***unitary*** if S is an invertible isometry.

Remark. Every isometry is injective, and every injective operator on a finite-dimensional vector space is invertible. Hence if V is finite-dimensional, we could delete the word "invertible" from the definition above without changing the meaning.

The unnecessary word "invertible" has been retained in the definition above for consistency with the definition readers may encounter when learning about inner product spaces that are not necessarily finite-dimensional.

Example (Rotation of $\mathbb{R}^2$). Suppose $\theta \in \mathbb{R}$. Let $S \in \mathscr{L}(\mathbf{F}^2)$ whose matrix with respect to the standard basis of $\mathbf{F}^2$ is

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

The two columns of this matrix form an orthonormal set in $\mathbf{F}^2$; hence S is an isometry, by 8.27. Thus S is a unitary operator.
If $\mathbf{F} = \mathbb{R}$, then S is the operator of counterclockwise rotation by θ radians around the origin of $\mathbb{R}^2$. This observation gives us another way to think about why S is an isometry, because each rotation around the origin of $\mathbb{R}^2$ preserves norms.

Lemma 8.29 (Characterisation of unitary operators). *Suppose $S \in \mathscr{L}(V)$, $\{e_1, \ldots, e_n\}$ is an orthonormal basis of V. Then the following are equivalent:*

(i) *S is a unitary operator.*

(ii) *$S^*S = SS^* = I$.*

(iii) *S is invertible, and $S^{-1} = S^*$.*

(iv) *$\{Se_1, \ldots, Se_n\}$ is an orthonormal basis of V.*

(v) *The rows of $\mathscr{M}(S; \{e_1, \ldots, e_n\})$ form an orthonormal basis of $\mathbf{F}^n$ with respect to the Euclidean inner product.*

(vi) *S^* is a unitary operator.*

Proof.

(i) $\Longrightarrow$ (ii) Suppose S is a unitary operator. Then S is an isometry, so $S^*S = I$ (by 8.27). Multiplying both sides by S^{-1} on the right yields $S^* = S^{-1}$. Thus $SS^* = SS^{-1} = I$ as desired.

(ii) $\Longrightarrow$ (iii) This follows from the definitions of invertibility and inverse.

(iii) $\Longrightarrow$ (iv) Suppose S is invertible and $S^{-1} = S^*$. Thus $S^*S = I$. By 8.27, $\{Se_1, \dots, Se_n\}$ is an orthonormal set in V.

Since this set has length $\dim V$, we conclude that $\{Se_1, \dots, Se_n\}$ is an orthonormal basis of V.

(iv) $\Longrightarrow$ (v) Suppose $\{Se_1, \dots, Se_n\}$ is an orthonormal basis of V. By 8.27, S is an isometry and thus a unitary operator. Hence

$$(S^*)^*S = SS^* = I$$

where the last equation holds since S^* is an isometry.

Thus the columns of $\mathscr{M}(S^*; \{e_1, \dots, e_n\})$ form an orthonormal basis of $\mathbf{F}^n$ (by 8.27). The rows of $\mathscr{M}(S; \{e_1, \dots, e_n\})$ are the complex conjugates of the columns of $\mathscr{M}(S^*; \{e_1, \dots, e_n\})$, so they form an orthonormal basis of $\mathbf{F}^n$.

(v) $\Longrightarrow$ (vi) Suppose the rows of $\mathscr{M}(S; \{e_1, \dots, e_n\})$ form an orthonormal basis of $\mathbf{F}^n$.

Thus the columns of $\mathscr{M}(S^*; \{e_1, \dots, e_n\})$ form an orthonormal basis of $\mathbf{F}^n$. By 8.27, S^* is an isometry and thus a unitary operator.

(vi) $\Longrightarrow$ (i) Suppose S^* is a unitary operator.

The chain of implications we have proven shows that (i) $\Longrightarrow$ (vi). Applying this result to S^* shows that $(S^*)^* = S$ is a unitary operator. □

Lemma 8.30. *Suppose λ is an eigenvalue of a unitary operator. Then $|\lambda| = 1$.*

Proof. Suppose $S \in \mathscr{L}(V)$ is a unitary operator. Let λ be an eigenvalue of S, with corresponding eigenvector v. Then

$$|\lambda| \|v\| = \|\lambda v\| = \|Sv\| = \|v\|.$$

Thus $|\lambda| = 1$, as desired. □

The next result characterises unitary operators on finite-dimensional complex inner product spaces, using the complex spectral theorem as the main tool.

Proposition 8.31. *Suppose $\mathbf{F} = \mathbb{C}$ and $S \in \mathscr{L}(V)$. Then S is a unitary operator if and only if there exists an orthonormal basis of V consisting of eigenvectors of S whose corresponding eigenvalues all have absolute value 1.*

Proof.

$\Longrightarrow$ Suppose S is a unitary operator.

By 8.29, $S^*S = SS^* = I$. Since S commutes with its adjoint, S is normal.

By the complex spectral theorem (8.20), there exists an orthonormal basis $\{e_1, \dots, e_n\}$ of V consisting of eigenvectors of S. By 8.30, every eigenvalue has absolute value 1.

$\boxed{\Longleftarrow}$ Let $\{e_1,\dots,e_n\}$ be an orthonormal basis of V consisting of eigenvectors of S, whose corresponding eigenvalues $\lambda_1,\dots,\lambda_n$ all have absolute value 1.

Then $\{Se_1,\dots,Se_n\}$ is an orthonormal basis of V, because

$$\begin{aligned}\langle Se_i, Se_j\rangle &= \langle \lambda_i e_i, \lambda_j e_j\rangle \\ &= \lambda_i\overline{\lambda_j}\langle e_i, e_j\rangle \\ &= \begin{cases} 0 & (i\neq j)\\ 1 & (i=j)\end{cases}\end{aligned}$$

for all $i,j\in\{1,\dots,n\}$. By 8.29, S is a unitary operator. □

QR Factorisation

We begin by making the following definition, transferring the notion of a unitary operator to a unitary matrix.

Definition 8.32 (Unitary matrix)**.** An $n\times n$ matrix is called ***unitary*** if its columns are orthonormal in $\mathbf{F}^n$.

Remark. In the definition above, we could have replaced "orthonormal set in $\mathbf{F}^n$" with "orthonormal basis of $\mathbf{F}^n$", because every orthonormal set of length n in an n-dimensional inner product space is an orthonormal basis.

Lemma 8.33 (Characterisation of unitary matrices)**.** *Suppose $Q\in\mathscr{M}_{n\times n}(\mathbf{F})$. Then the following are equivalent:*

(i) *Q is a unitary matrix.*

(ii) *The rows of Q form an orthonormal set in $\mathbf{F}^n$.*

(iii) *$\|Qv\| = \|v\|$ for every $v\in\mathbf{F}^n$.*

(iv) *$Q^*Q = QQ^* = I_n$.*

Proof. □

The QR factorisation is the main tool in the widely used QR algorithm (not discussed here) for finding good approximations to eigenvalues and eigenvectors of square matrices.

Theorem 8.34 (QR factorisation)**.** *Suppose A is a square matrix with linearly independent columns. Then there exist unique unitary matrix Q, and upper-triangular matrix R*

with only positive numbers on its diagonal, such that

$$A = QR. \tag{8.1}$$

Proof.

Existence Let $v_1, \ldots, v_n$ denote the columns of A (consider these as elements of $\mathbf{F}^n$). Apply the Gram–Schmidt procedure to the linearly independent set $\{v_1, \ldots, v_n\}$ to obtain an orthonormal basis of $\{e_1, \ldots, e_n\}$ of $\mathbf{F}^n$ such that

$$\operatorname{span}(v_1, \ldots, v_i) = \operatorname{span}(e_1, \ldots, e_i)$$

for each $i = 1, \ldots, n$.

Claim. Let Q be the matrix whose columns are $e_1, \ldots, e_n$.

Thus Q is unitary.

Claim. Let R be the $n \times n$ matrix defined by $R_{ij} = \langle v_j, e_i \rangle$.

If $i > j$, then e_i is orthogonal to $\operatorname{span}(e_1, \ldots, e_j)$ and hence e_i is orthogonal to v_j (by 7.59); thus $R_{ij} = \langle v_j, e_i \rangle = 0$, so R is upper-triangular.

The equations defining the Gram–Schmidt procedure (see 6.32) show that each v_i equals a positive multiple of e_i plus a linear combination of $e_1, \ldots, e_{i-1}$. Thus each $\langle v_i, e_i \rangle$ is a positive number. Hence all entries on the diagonal of R are positive numbers.

If $i \in \{1, \ldots, n\}$, then the i-th column of QR equals a linear combination of the columns of Q, with the coefficients for the linear combination coming from the i-th column of R. Hence the i-th column of QR equals

$$\langle v_i, e_1 \rangle e_1 + \cdots + \langle v_i, e_i \rangle e_i$$

which equals v_i, the i-th column of A. Hence $A = QR$, as desired.

Uniqueness Suppose we also have $A = \hat{Q}\hat{R}$, where $\hat{Q}$ is unitary and $\hat{R}$ is upper-triangular with only positive numbers on its diagonal. □

Cholesky Factorisation

We begin this subsection with a characterisation of positive invertible operators in terms of inner products.

Lemma 8.35 (Positive invertible operator). *A self-adjoint operator $T \in \mathscr{L}(V)$ is a positive invertible operator if and only if $\langle Tv, v \rangle > 0$ for every non-zero $v \in V$.*

Proof.

$\boxed{\Longrightarrow}$ Suppose T is a positive invertible operator.

Let $v \in V \setminus \{\mathbf{0}\}$. Since T is invertible, T is injective so $Tv \neq \mathbf{0}$. This implies that $\langle Tv, v\rangle \neq 0$ (by 7.43). Hence $\langle Tv, v\rangle > 0$.

$\boxed{\Longleftarrow}$ Suppose $\langle Tv, v\rangle > 0$ for every $v \in V \setminus \{\mathbf{0}\}$. Thus $Tv \neq \mathbf{0}$ for every $v \in V \setminus \{\mathbf{0}\}$.

Hence T is injective, and thus is invertible. □

Definition 8.36 (Positive definite). A matrix $B \in \mathscr{M}_{n\times n}(\mathbf{F})$ is called ***positive definite*** if

(i) B is Hermitian ($B^* = B$), and

(ii) $\langle Bx, x\rangle > 0$ for every non-zero $x \in \mathbf{F}^n$.

A matrix is upper triangular if and only if its conjugate transpose is lower triangular (meaning that all entries above the diagonal are 0). The factorisation below writes a positive definite matrix as the product of a lower triangular matrix and its conjugate transpose.

Theorem 8.37 (Cholesky factorisation). *Suppose B is a positive definite matrix. Then there exists a unique upper-triangular matrix R with only positive numbers on its diagonal such that*

$$B = R^*R. \tag{8.2}$$

Proof. □

8.5 Singular Value Decomposition

Singular Values

We will need the following result in this section.

Lemma 8.38 (Properties of T^*T). *Suppose $T \in \mathscr{L}(V,W)$. Then*

(i) *T^*T is a positive operator on V;*

(ii) $\ker T^*T = \ker T$;

(iii) $\operatorname{im} T^*T = \operatorname{im} T$;

(iv) $\dim \operatorname{im} T = \dim \operatorname{im} T^* = \dim \operatorname{im} T^*T$.

Proof.

(i) Since

$$(T^*T)^* = T^*(T^*)^* = T^*T,$$

T^*T is self-adjoint.

Let $v \in V$. Then

$$\langle (T^*T)v, v\rangle = \langle T^*(Tv), v\rangle = \langle Tv, Tv\rangle = \|Tv\|^2 \geq 0.$$

Thus T^*T is a positive operator.

(ii) $\boxed{\subset}$ Let $v \in \ker T^*T$. Then

$$\|Tv\|^2 = \langle Tv, Tv\rangle = \langle T^*Tv, v\rangle = \langle \mathbf{0}, v\rangle = 0.$$

Thus $Tv = \mathbf{0}$. Hence $\ker T^*T \subset \ker T$.

$\boxed{\supset}$ Let $v \in \ker T$. Then $Tv = \mathbf{0}$, so $T^*Tv = \mathbf{0}$. Hence $\ker T \subset \ker T^*T$.

(iii) By (i), T^*T is self-adjoint. Thus

$$\operatorname{im} T^*T = (\ker T^*T)^\perp = (\ker T)^\perp = \ker T^*.$$

(iv) For the first equality,

$$\dim \operatorname{im} T = \dim(\ker T^*)^\perp = \dim W - \dim \ker T^* = \dim \operatorname{im} T^*.$$

The second equality $\dim \operatorname{im} T^* = \dim \operatorname{im} T^*T$ follows from (iii).

□

Definition 8.39 (Singular values). Suppose $T \in \mathscr{L}(V,W)$. The ***singular values*** of T are the non-negative square roots of the eigenvalues of T^*T

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0,$$

each included as many times as the dimension of the corresponding eigenspace of T^*T.

Lemma 8.40 (Role of positive singular values). *Suppose $T \in \mathscr{L}(V,W)$. Then*

(i) T is injective $\iff$ 0 is not a singular value of T;

(ii) the number of positive singular values of T equals $\dim \operatorname{im} T$*;*

(iii) T is surjective $\iff$ number of positive singular values of T equals $\dim W$.

Proof.

(i) We have

$$\begin{aligned} T \text{ is injective} &\iff \ker T = \{\mathbf{0}\} \\ &\iff \ker T^*T = \{\mathbf{0}\} \\ &\iff 0 \text{ is not an eigenvalue of } T^*T \\ &\iff 0 \text{ is not a singular value of } T. \end{aligned}$$

(ii) The spectral theorem applied to T^*T shows that $\dim \operatorname{im} T^*T$ equals the number of positive eigenvalues of T^*T (counting repetitions). Thus 7.64(d) implies that $\dim \operatorname{im} T$ equals the number of positive singular values of T.

(iii) This follows from (ii) and 2.39.

□

The next result characterises isometries in terms of singular values.

Lemma 8.41. *Suppose $S \in \mathscr{L}(V,W)$. Then*

$$S \text{ is an isometry} \iff \text{all singular values of } S \text{ equal } 1.$$

Proof. We have

$$\begin{aligned} S \text{ is an isometry} &\iff S^*S = I \\ &\iff \text{all eigenvalues of } S^*S \text{ equal } 1 \\ &\iff \text{all singular values of } S \text{ equal } 1 \end{aligned}$$

□

SVD for Linear Maps and for Matrices

The next result shows that every linear map from V to W has a remarkably clean description in terms of its singular values and orthonormal sets in V and W; this is known as *singular value decomposition* (SVD).

Theorem 8.42 (Singular value decomposition). *Suppose $T \in \mathscr{L}(V,W)$ and the positive singular values of T are $\sigma_1,\dots,\sigma_m$. Then there exist orthonormal sets $\{e_1,\dots,e_m\}$ in V and $\{f_1,\dots,f_m\}$ in W such that*

$$Tv = \sigma_1 \langle v, e_1\rangle f_1 + \cdots + \sigma_m \langle v, e_m\rangle f_m \tag{8.3}$$

for every $v \in V$.

Proof. Let $\sigma_1,\dots,\sigma_n$ denote the singular values of T (thus $n = \dim V$).

1. Since T^*T is a positive operator, by the spectral theorem, there exists an orthonormal basis $\{e_1,\dots,e_n\}$ of V with
$$T^*Te_i = \sigma_i e_i^2 \quad (i = 1,\dots,n).$$

2. For each $i = 1,\dots,m$, let
$$f_i = \frac{Te_i}{\sigma_i}.$$
We check that $\{f_1,\dots,f_m\}$ is an orthonormal set in W. If $i,j \in \{1,\dots,m\}$, then
$$\begin{aligned}\langle f_i, f_j\rangle &= \left\langle \frac{1}{\sigma_i}Te_i, \frac{1}{\sigma_j}Te_j \right\rangle \\ &= \frac{1}{\sigma_i\sigma_j}\langle Te_i, Te_j\rangle \\ &= \frac{1}{\sigma_i\sigma_j}\langle e_i, T^*Te_j\rangle \\ &= \frac{\sigma_j}{\sigma_i}\langle e_i, e_j\rangle \\ &= \begin{cases} 0 & (i \neq j) \\ 1 & (i = j) \end{cases}\end{aligned}$$

3. If $i \in \{1,\dots,n\}$ and $i > m$, then $\sigma_i = 0$ and hence $T^*Te_i = 0$, which implies that $Te_i = 0$.

Let $v \in V$. Then

$$\begin{aligned} Tv &= T\left(\langle v, e_1\rangle e_1 + \cdots + \langle v, e_n\rangle e_n\right) \\ &= \langle v, e_1\rangle Te_1 + \cdots + \langle v, e_m\rangle Te_m \\ &= \sigma_1 \langle v, e_1\rangle f_1 + \cdots + \sigma_m \langle v, e_m\rangle f_m. \end{aligned}$$

□

Suppose $T \in \mathscr{T}(V,W)$, with singular values $\sigma_1, \ldots, \sigma_m$. Let $\{e_1, \ldots, e_m\}$ and $\{f_1, \ldots, f_m\}$ be such that (8.3) holds. Extend the orthonormal set $\{e_1, \ldots, e_m\}$ to an orthonormal basis $\{e_1, \ldots, e_{\dim V}\}$ of V, and extend the orthonormal set $\{f_1, \ldots, f_m\}$ to an orthonormal basis $\{f_1, \ldots, f_{\dim W}\}$ of W. (8.3) shows that

$$Te_i = \begin{cases} \sigma_i f_i & (1 \le i \le m) \\ 0 & (m < i \le \dim V) \end{cases}$$

Thus the matrix of T with respect to the orthonormal bases $\{e_1, \ldots, e_{\dim V}\}$ and $\{f_1, \ldots, f_{\dim W}\}$ is

$$\mathscr{M}(T)_{ij} = \begin{cases} \sigma_i & (1 \le i = j \le m) \\ 0 & (\text{otherwise}) \end{cases}$$

If $\dim V = \dim W$ (when $V = W$), then the matrix described in the paragraph above is a diagonal matrix. Let us extend the definition of diagonal matrix to matrices that are not necessarily square:

An $M \times N$ matrix A is called a *diagonal matrix* if all entries of the matrix are 0 except possibly A_{ii} for $i = 1, \ldots, \min\{M, N\}$.

Then we have shown that *every linear map has a diagonal matrix with respect to some orthonormal bases*.

Theorem 8.43 (Singular value decomposition of adjoint and pseudoinverse)**.** *Suppose $T \in \mathscr{L}(V,W)$ with singular values $\sigma_1, \ldots, \sigma_m$. Suppose $\{e_1, \ldots, e_m\}$ and $\{f_1, \ldots, f_m\}$ are orthonormal sets in V and W such that*

$$Tv = \sigma_1 \langle v, e_1\rangle f_1 + \cdots + \sigma_m \langle v, e_m\rangle f_m$$

for every $v \in V$. Then

$$T^* w = \sigma_1 \langle w, f_1\rangle e_1 + \cdots + \sigma_m \langle w, f_m\rangle e_m \tag{8.4}$$

and

$$T^+w = \frac{\langle w, f_1\rangle}{\sigma_1} e_1 + \cdots + \frac{\langle w, f_m\rangle}{\sigma_m} e_m \tag{8.5}$$

for every $w \in W$.

Proof.

Adjoint Let $v \in V$, $w \in W$. Then

$$\begin{aligned}\langle Tv, w\rangle &= \langle \sigma_1 \langle v, e_1\rangle f_1 + \cdots + \sigma_m \langle v, e_m\rangle f_m, w\rangle \\ &= \sigma_1 \langle v, e_1\rangle \langle f_1, w\rangle + \cdots + \sigma_m \langle v, e_m\rangle \langle f_m, w\rangle \\ &= \langle v, \sigma_1 \langle w, f_1\rangle e_1 + \cdots + \sigma_m \langle w, f_m\rangle e_m\rangle .\end{aligned}$$

Thus (8.4) follows.

Pseudoinverse Let $w \in W$. Let

$$v = \frac{\langle w, f_1\rangle}{\sigma_1} e_1 + \cdots + \frac{\langle w, f_m\rangle}{\sigma_m} e_m.$$

Applying T to both sides gives

$$\begin{aligned}Tv &= \frac{\langle w, f_1\rangle}{\sigma_1} Te_1 + \cdots + \frac{\langle w, f_m\rangle}{\sigma_m} Te_m \\ &= \langle w, f_1\rangle f_1 + \cdots + \langle w, f_m\rangle f_m \\ &= P_{\operatorname{im} T} w.\end{aligned}$$

... Thus $v = T^+w$, and (8.5) follows. □

Theorem 8.44 (Singular value decomposition, matrix). *Suppose* $A \in \mathscr{M}_{p\times n}(\mathbf{F})$ *has rank* $m \geq 1$. *Then*

$$A = U\Sigma V^* \tag{8.6}$$

for some $U \in \mathscr{M}_{p\times m}(\mathbf{F})$ *with orthonormal columns,* $\Sigma \in \mathscr{M}_{m\times m}(\mathbf{F})$ *with positive numbers on the diagonal,* $V \in \mathscr{M}_{n\times m}(\mathbf{F})$ *with orthonormal columns.*

Proof. Let $T : \mathbf{F}^n \to \mathbf{F}^p$ be the linear map whose matrix with respect to the standard bases equals A. Then $\dim \operatorname{im} T = m$. Let

$$Tv = \sigma_1 \langle v, e_1\rangle f_1 + \cdots + \sigma_m \langle v, e_m\rangle f_m$$

be a singular value decomposition of T.

Claim. Let

- U be the $p \times m$ matrix whose columns are $f_1, \ldots, f_m$,

- Σ be the $m \times m$ diagonal matrix whose diagonal entries are $\sigma_1, \ldots, \sigma_m$,
- V be the $n \times m$ matrix whose columns are $e_1, \ldots, e_m$.

We now show (8.6) holds. Let $\{u_1, \ldots, u_m\}$ denote the standard basis of $\mathbf{F}^m$. For each $i = 1, \ldots, m$,

$$\begin{aligned}(AV - U\Sigma)u_i &= A(Vu_i) - U(\Sigma u_i) \\ &= Ae_i - U(\sigma_i u_i) \\ &= \sigma_i f_i - \sigma_i f_i = 0\end{aligned}$$

implies $AV = U\Sigma$. Then multiply both sides by V^* on the right to get

$$AVV^* = U\Sigma V^*.$$

Claim. $AVV^* = A$.

Note that the rows of V^* are the complex conjugates of $e_1, \ldots, e_m$. Thus if $i \in \{1, \ldots, m\}$, then the definition of matrix multiplication shows that $V^* e_i = u_i$; hence $VV^* e_i = e_i$. Thus $AVV^* v = Av$ for all $v \in \operatorname{span}(e_1, \ldots, e_m)$.

If $v \in (\operatorname{span}(e_1, \ldots, e_m))^\perp$, then $Av = 0$ (as follows from 7.81) and $V^* v = 0$ (as follows from the definition of matrix multiplication). Hence $AVV^* v = Av$ for all $v \in (\operatorname{span}(e_1, \ldots, e_m))^\perp$.

Since AVV^* and A agree on $\operatorname{span}(e_1, \ldots, e_m)$ and on $(\operatorname{span}(e_1, \ldots, e_m))^\perp$, we conclude that $AVV^* = A$. Thus (8.6) follows. □

8.6 Consequences of Singular Value Decomposition

Norm of Linear Map

The singular value decomposition leads to an upper bound for $\|Tv\|$.

Lemma 8.45. *Suppose $T \in \mathscr{L}(V,W)$. Then $\|Tv\| \leq \sigma_1 \|v\|$ for all $v \in V$.*

Proof. Suppose $T \in \mathscr{L}(V,W)$, with singular values $\sigma_1, \ldots, \sigma_m$. Let $\{e_1, \ldots, e_m\}$ be an orthonormal set in V, and $\{f_1, \ldots, f_m\}$ be an orthonormal set in W that provide a singular value decomposition of T. Thus

$$Tv = \sigma_1 \langle v, e_1 \rangle f_1 + \cdots + \sigma_m \langle v, e_m \rangle f_m$$

for all $v \in V$. Then

$$\begin{aligned} \|Tv\|^2 &= {\sigma_1}^2 |\langle v, e_1 \rangle|^2 + \cdots + {\sigma_m}^2 |\langle v, e_m \rangle|^2 \\ &\leq {\sigma_1}^2 \left(|\langle v, e_1 \rangle|^2 + \cdots + |\langle v, e_m \rangle|^2 \right) \\ &\leq {\sigma_1}^2 \|v\|^2, \end{aligned}$$

where the last inequality follows from Bessel's inequality. Taking square roots on both sides yields the desired inequality. □

Definition 8.46 (Norm of linear map)**.** Suppose $T \in \mathscr{L}(V,W)$. Then the ***norm*** of T is

$$\|T\| := \sup_{\|v\| \leq 1} \|Tv\|.$$

Lemma 8.47 (Basic properties of norm of linear map)**.** *Suppose $T \in \mathscr{L}(V,W)$. Then*

(i) $\|T\| \geq 0$*, where equality holds if and only if* $T = 0$*;* *(positive definiteness)*

(ii) $\|\lambda T\| = |\lambda| \|T\|$ *for all* $\lambda \in \mathbf{F}$*;* *(homogeneity)*

(iii) $\|S+T\| \leq \|S\| + \|T\|$ *for all* $S \in \mathscr{L}(V,W)$*.* *(triangle inequality)*

Proof.

(i) Since $\|Tv\| \geq 0$ for every $v \in V$, the definition of $\|T\|$ implies $\|T\| \geq 0$.

(ii) $\boxed{\Longrightarrow}$ Suppose $\|T\| = 0$. Thus $Tv = \mathbf{0}$ for all $v \in V$, $\|v\| \leq 1$.

Let $u \in V$, $u \neq \mathbf{0}$. Then

$$Tu = \|u\| T\left(\frac{u}{\|u\|}\right) = \mathbf{0},$$

where the last equality holds since $u/\|u\|$ has norm 1. Since $Tu = \mathbf{0}$ for all $u \in V$, we have $T = 0$.

$\boxed{\Longleftarrow}$ If $T = 0$, then $Tv = \mathbf{0}$ for all $v \in V$. Hence $\|T\| = 0$.

(iii) Let $\lambda \in \mathbf{F}$. Then

$$\|\lambda T\| = \sup_{\|v\| \leq 1} \|\lambda T v\| = |\lambda| \sup_{\|v\| \leq 1} \|Tv\| = |\lambda| \|T\|.$$

(iv) Let $S \in \mathscr{L}(V,W)$. The definition of $\|S+T\|$ implies there exists $v \in V$, $\|v\| \leq 1$ such that $\|S+T\| = \|(S+T)v\|$. Then

$$\|S+T\| = \|(S+T)v\| = \|Sv+Tv\| \leq \|Sv\| + \|Tv\| \leq \|S\| + \|T\|.$$

□

Hence $\mathscr{L}(V,W)$ is a metric space, with metric $d(S,T) = \|S-T\|$ for $S,T \in \mathscr{L}(V,W)$.

(i) $d(S,S) = \|S-S\| = 0$. If $S \neq T$, then $d(S,T) = \|S-T\| = \sigma_{\max}(S-T)$. Since $S-T \neq 0$, its largest singular value is nonzero and therefore $d(S,T) > 0$.

(ii) $d(S,T) = \|S-T\| = \|T-S\| = d(T,S)$.

(iii) $d(S,G) = \|S-G\| \leq \|S-T\| + \|T-G\| = d(S,T) + d(T,G)$.

Lemma 8.48 (Alternative formulae for $\|T\|$). *Suppose $T \in \mathscr{L}(V,W)$. Then*

(i) $\|T\|$ = *largest singular value of* T;

(ii) $\|T\| = \sup_{\|v\|=1} \|Tv\|$;

(iii) $\|T\|$ = *smallest number* c *such that* $\|Tv\| \leq c\|v\|$ *for all* $v \in V$.

Proof.

(i) This follows from 8.45.

(ii) Let $v \in V$, $\|v\| \leq 1$. Let $u = v/\|v\|$. Then

$$\|u\| = \left\| \frac{v}{\|v\|} \right\|$$

and

$$\|Tu\| = \left\| T\left(\frac{v}{\|v\|}\right)\right\| = \frac{\|Tv\|}{\|v\|} \geq \|Tv\|.$$

Thus when finding the maximum of $\|Tv\|$ with $\|v\| \leq 1$, we can restrict our attention to vectors in V with norm 1, proving (ii).

(iii) Let $v \in V$, $v \neq \mathbf{0}$. Then the definition of $\|T\|$ implies that

$$\left\| T\left(\frac{v}{\|v\|}\right)\right\| \leq \|T\|,$$

which implies that

$$\|Tv\| \leq \|T\|\|v\|.$$

Now suppose $c \geq 0$ and $\|Tv\| \leq c\|v\|$ for all $v \in V$. This implies that

$$\|Tv\| \leq c$$

for all $v \in V$, $\|v\| \leq 1$. Taking sup on the LHS over all $v \in V$, $\|v\| \leq 1$ shows that $\|T\| \leq c$. Thus $\|T\|$ is the smallest number c such that $\|Tv\| \leq c\|v\|$ for all $v \in V$.

□

An important inequality during the proof is

$$\|Tv\| \leq \|T\|\|v\| \tag{8.7}$$

for all $v \in V$, $v \neq \mathbf{0}$.

A linear map and its adjoint have the same norm, as shown by the next result.

Lemma 8.49 (Norm of adjoint). *Suppose $T \in \mathscr{L}(V,W)$. Then $\|T^*\| = \|T\|$.*

Proof. Suppose $w \in W$. Then

$$\|T^*w\|^2 = \langle T^*w, T^*w\rangle = \langle TT^*w, w\rangle \leq \|TT^*w\|\|w\| \leq \|T\|\|T^*w\|\|w\|.$$

The inequality above implies that

$$\|T^*w\| \leq \|T\|\|w\|.$$

But $\|T^*w\| \leq \|T^*\|\|w\|$, so we have $\|T^*\| \leq \|T\|$.

Replacing T with T^* shows that $\|T\| \leq \|T^*\|$. Thus $\|T^*\| = \|T\|$, as desired. □

Approximation by Linear Maps with Lower-Dimensional Range

Theorem 8.50 (Best approximation by linear map whose image has dimension $\leq k$)**.** *Suppose $T \in \mathscr{L}(V,W)$, and $\sigma_1, \ldots, \sigma_m$ are the singular values of T. Suppose $1 \leq k < m$. Then*

$$\min\{\|T - S\| \mid S \in \mathscr{L}(V,W) \text{ and } \dim \operatorname{im} S \leq k\} = \sigma_{k+1}. \tag{8.8}$$

Furthermore, if

$$Tv = \sigma_1 \langle v, e_1 \rangle f_1 + \cdots + \sigma_m \langle v, e_m \rangle f_m$$

is a singular value decomposition of T and $T_k \in \mathscr{L}(V,W)$ is defined by

$$T_k v = \sigma_1 \langle v, e_1 \rangle f_1 + \cdots + \sigma_k \langle v, e_k \rangle f_k$$

for each $v \in V$, then $\dim \operatorname{im} T_k = k$ and $\|T - T_k\| = \sigma_{k+1}$.

Polar Decomposition

Every non-zero complex number $z \in \mathbb{C}$ can be written in the form

$$\begin{aligned} z &= \left(\frac{z}{|z|}\right)|z| \\ &= \left(\frac{z}{|z|}\right)\sqrt{\bar{z}z} \end{aligned}$$

where $z/|z|$ has absolute value 1, and $\sqrt{\bar{z}z}$ is positive.

Our analogy leads us to guess that every operator $T \in \mathscr{L}(V)$ can be written as a unitary operator times $\sqrt{T^*T}$. The corresponding result is called the *polar decomposition*, which gives a beautiful description of an arbitrary operator on V.

Theorem 8.51 (Polar decomposition)**.** *Suppose $T \in \mathscr{L}(V)$. Then there exists a unitary operator $S \in \mathscr{L}(V)$ such that*

$$T = S\sqrt{T^*T}. \tag{8.9}$$

Remark. This holds for both $\mathbb{C}$ and $\mathbb{R}$.

Proof. Let $\sigma_1, \ldots, \sigma_m$ be the positive singular values of T. Let $\{e_1, \ldots, e_m\}$ and $\{f_1, \ldots, f_m\}$ be orthonormal sets in V such that

$$Tv = \sigma_1 \langle v, e_1 \rangle f_1 + \cdots + \sigma_m \langle v, e_m \rangle f_m$$

for every $v \in V$. Extend $\{e_1, \ldots, e_m\}$ and $\{f_1, \ldots, f_m\}$ to orthonormal bases $\{e_1, \ldots, e_n\}$ and $\{f_1, \ldots, f_n\}$ of V.

Recall that the singular value decomposition of the adjoint is

$$T^*w = \sigma_1 \langle w, f_1 \rangle e_1 + \cdots + \sigma_m \langle w, f_m \rangle e_m$$

for all $w \in W$. Thus

$$T^*Tv = \sigma_1^2 \langle v, e_1 \rangle e_1 + \cdots + \sigma_m^2 \langle v, e_m \rangle e_m$$

for every $v \in V$. Then

$$\sqrt{T^*T}v = \sigma_1 \langle v, e_1 \rangle e_1 + \cdots + \sigma_m \langle v, e_m \rangle e_m$$

because the operator that sends v to the RHS of the equation above is a positive operator whose square equals T^*T.

Claim. Define $S \in \mathscr{L}(V)$ by

$$Sv = \langle v, e_1 \rangle f_1 + \cdots + \langle v, e_n \rangle f_n$$

for each $v \in V$.

Then

$$\begin{aligned} \|Sv\|^2 &= \|\langle v, e_1 \rangle f_1 + \cdots + \langle v, e_n \rangle f_n\|^2 \\ &= |\langle v, e_1 \rangle|^2 + \cdots + |\langle v, e_n \rangle|^2 \\ &= \|v\|^2. \end{aligned}$$

Thus S is a unitary operator. Now

$$\begin{aligned} S\sqrt{T^*T}v &= S(\sigma_1 \langle v, e_1 \rangle e_1 + \cdots + \sigma_m \langle v, e_m \rangle e_m) \\ &= \sigma_1 \langle v, e_1 \rangle f_1 + \cdots + \sigma_m \langle v, e_m \rangle f_m \\ &= Tv. \end{aligned}$$

□

Operators Applied to Ellipsoids and Parallelepipeds

Definition 8.52 (Ball). The ***unit ball*** in V centred at $\mathbf{0}$ is

$$B := \{v \in V \mid \|v\| \leq 1\}.$$

You can think of the ellipsoid defined below as obtained by starting with the ball B, and then stretching by a factor of s_i along each f_i-axis.

Definition 8.53 (Ellipsoid). Suppose $\{f_1,\dots,f_n\}$ is an orthonormal basis of V, and $s_1,\dots,s_n > 0$. The ***ellipsoid*** with principal axes $s_1f_1,\dots,s_nf_n$ is

$$E(s_1f_1,\dots,s_nf_n) := \left\{ v \in V \;\middle|\; \frac{|\langle v,f_1\rangle|^2}{s_1} + \cdots + \frac{|\langle v,f_n\rangle|^2}{s_n} < 1 \right\}.$$

Remark. If $\dim V = 2$, the word "disk" is sometimes used to denote ball and the word "ellipse" is sometimes used to denote ellipsoid.

The next result states that an invertible map takes a ball to an ellipsoid.

Proposition 8.54. *Suppose $T \in \mathscr{L}(V)$ is invertible. Then T maps the ball B in V to an ellipsoid in V.*

Proof. Suppose T has the singular value decomposition

$$Tv = \sigma_1 \langle v,e_1\rangle f_1 + \cdots + \sigma_n \langle v,e_n\rangle f_n$$

for all $v \in V$, where $\{e_1,\dots,e_n\}$ and $\{f_1,\dots,f_n\}$ are orthonormal bases of V. We will show that

$$T(B) = E(\sigma_1 f_1,\dots,\sigma_n f_n).$$

$\boxed{\subset}$ Let $v \in B$. Since T is invertible, none of the singular values $\sigma_1,\dots,\sigma_n$ equal 0. Thus

$\boxed{\supset}$ □

The next result states that an invertible map takes an ellipsoid to an ellipsoid.

Proposition 8.55. *Suppose $T \in \mathscr{L}(V)$ is invertible, and E is an ellipsoid in V. Then $T(E)$ is an ellipsoid in V.*

Definition 8.56 (Parallelepiped). Suppose $\{v_1,\dots,v_n\}$ is a basis of V. Let

$$P(v_1,\dots,v_n) := \{a_1v_1 + \cdots + a_nv_n \mid a_i \in (0,1)\}.$$

A ***parallelepiped*** is a set of the form $u + P(v_1,\dots,v_n)$ for some $u \in V$. The vectors $v_1,\dots,v_n$ are called the *edges* of the parallelepiped.

The next result states an invertible operator takes a parallelepiped to a parallelepiped.

Proposition 8.57. *Suppose $u \in V$, and $\{v_1,\dots,v_n\}$ is a basis of V. Suppose $T \in \mathscr{L}(V)$ is invertible. Then*

$$T(u + P(v_1,\dots,v_n)) = Tu + P(Tv_1,\dots,Tv_n).$$

Definition 8.58 (Box)**.** A ***box*** is of the form

$$u + P(r_1 e_1, \ldots, r_n e_n)$$

where $u \in V$, $r_1, \ldots, r_n > 0$, and $\{e_1, \ldots, e_n\}$ is an orthonormal basis of V.

Volume via Singular Values

Properties of an Operator as Determined by Its Eigenvalues

Exercises

7A 1-12 15 16 17 18

Exercise 8.1 ([Axl24] 7A Q12)**.** An operator $B \in \mathscr{L}(V)$ is called *skew* if

$$B^* = -B.$$

Suppose $T \in \mathscr{L}(V)$. Prove that T is normal if and only if there exist commuting operators A and B such that A is self-adjoint, B is a skew operator, and $T = A + B$.

Solution. □

Exercise 8.2 ([Axl24] 7A Q19)**.** Suppose $T \in \mathscr{L}(V)$ and $\|T^*v\| \leq \|Tv\|$ for every $v \in V$. Prove that T is normal.

Remark. This exercise fails on infinite-dimensional inner product spaces, leading to what are called *hyponormal operators*.

Solution. Let $\{e_1, \ldots, e_n\}$ be an orthonormal basis of V. For $i = 1, \ldots, n$, write

$$\begin{aligned} \|Te_i\|^2 &= |\langle Te_i, e_1\rangle|^2 + \cdots + |\langle Te_i, e_n\rangle|^2 \\ &= |\langle e_i, T^*e_1\rangle|^2 + \cdots + |\langle e_i, T^*e_n\rangle|^2. \end{aligned}$$

Summing over i,

$$\begin{aligned} \sum_{i=1}^{n} \|Te_i\|^2 &= \sum_{i=1}^{n} |\langle e_i, T^*e_1\rangle|^2 + \cdots + \sum_{i=1}^{n} |\langle e_i, T^*e_n\rangle|^2 \\ &= \|T^*e_1\|^2 + \cdots + \|T^*e_n\|^2. \end{aligned}$$

Since we are given $\|T^*v\| \leq \|Tv\|$ for every $v \in V$, and equality holds, we must have $\|Te_i\| = \|T^*e_i\|$ for each $i = 1, \ldots, n$. Since the choice of orthonormal basis was arbitrary, we must have $\|Tu\| = \|T^*u\|$ for every unit vector $u \in V$.

For every $v \in V$, $\frac{1}{\|v\|}v \in V$ is a unit vector, so

$$\left\| T\left(\frac{1}{\|v\|}v\right) \right\| = \left\| T^*\left(\frac{1}{\|v\|}v\right) \right\|$$

which simplifies to $\|Tv\| = \|T^*v\|$. Hence by 8.13, T is normal. □

7A 20

Exercise 8.3 ([Axl24] 7A Q24)**.** Suppose $T \in \mathscr{L}(V)$ and

$$a_0 + a_1 z + a_2 z^2 + \cdots + a_{m-1} z^{m-1} + z^m$$

is the minimal polynomial of T. Prove that the minimal polynomial of T^* is

$$\overline{a_0} + \overline{a_1} z + \overline{a_2} z^2 + \cdots + \overline{a_{m-1}} z^{m-1} + z^m.$$

Remark. This exercise shows that the minimal polynomial of T^* equals the minimal polynomial of T if $\mathbf{F} = \mathbb{R}$.

Solution. Let p be the minimal polynomial of T.

Claim. If f is any polynomial, then $f(T^*) = \overline{f(T)}^*$.

Let $f(x) = c_n x^n + c_{n-1} x^{n-1} + \cdots + c_1 x + c_0$. Then

$$\begin{aligned} f(T^*) &= c_n (T^*)^n + c_{n-1} (T^*)^{n-1} + \cdots + c_1 T^* + c_0 \\ &= c_n (T^n)^* + \cdots + c_1 T^* + c_0 \\ &= (\overline{c_n} T^n + \cdots + \overline{c_1} T + \overline{c_0})^* \\ &= \overline{f(T)}^* \end{aligned}$$

as desired.

Since p is the minimal polynomial of T, we have $p(T) = 0$, so

$$\overline{p(T^*)} = p(T)^* = 0^* = 0$$

which implies $\overline{p}$ is a zero polynomial of T^*.

Let q be the minimal polynomial of T^*. Then $\overline{q}$ is the minimal polynomial of $(T^*)^* = T$. Since p is the minimal polynomial, $p \mid \overline{q}$ which implies $\overline{p} \mid q$. Hence $\overline{p} = q$ by minimality of q. □

Exercise 8.4 ([Axl24] 7A Q25)**.** Suppose $T \in \mathscr{L}(V)$. Prove that T is diagonalisable if and only if T^* is diagonalisable.

Solution. □

7A 27 28 29

Exercise 8.5 ([Axl24] 7B Q1)**.** Prove that a normal operator on a complex inner product space is self-adjoint if and only if all its eigenvalues are real.

Exercise 8.6 ([Axl24] 7B Q2)**.** Suppose $\mathbf{F} = \mathbb{C}$. Suppose $T \in \mathscr{L}(V)$ is normal and has only one eigenvalue. Prove that T is a scalar multiple of the identity operator.

Exercise 8.7 ([Axl24] 7B Q3)**.** Suppose $\mathbf{F} = \mathbb{C}$ and $T \in \mathscr{L}(V)$ is normal. Prove that the set of eigenvalues of T is contained in $\{0,1\}$ if and only if there is a subspace U of V such that $T = P_U$.

Exercise 8.8 ([Axl24] 7B Q4)**.** Prove that a normal operator on a complex inner product space is skew (meaning it equals the negative of its adjoint) if and only if all its eigenvalues are purely imaginary.

Exercise 8.9 ([Axl24] 7B Q6)**.** Suppose V is a complex inner product space and $T \in \mathscr{L}(V)$ is a normal operator such that $T^9 = T^8$. Prove that T is self-adjoint and $T^2 = T$.

Exercise 8.10 ([Axl24] 7B Q8)**.** Suppose $\mathbf{F} = \mathbb{C}$ and $T \in \mathscr{L}(V)$. Prove that T is normal if and only if every eigenvector of T is also an eigenvector of T^*.

Exercise 8.11 ([Axl24] 7B Q9)**.** Suppose $\mathbf{F} = \mathbb{C}$ and $T \in \mathscr{L}(V)$. Prove that T is normal if and only if there exists a polynomial $p \in \mathbb{C}[z]$ such that $T^* = p(T)$.

Solution.

$\boxed{\Longleftarrow}$ Suppose there exists a polynomial $p \in \mathbb{C}[z]$ such that $T^* = p(T)$. Since $Tp(T) = p(T)T$, this implies $TT^* = T^*T$ so T is normal.

$\boxed{\Longrightarrow}$ Let $\lambda_1, \dots, \lambda_r$ be eigenvalues of T. Then

$$V = \bigoplus_{i=1}^{r} E(\lambda_i, T).$$

Since T is normal, we have $E(\lambda_i, T) = E(\overline{\lambda_i}, T^*)$ for each i. Thus

$$V = \bigoplus_{i=1}^{r} E(\overline{\lambda_i}, T^*).$$

Let $V_i = E(\lambda_i, T) = E(\overline{\lambda_i}, T^*)$. Then

$$T|_{V_i} = \lambda_i I_{V_i}, \quad T^*|_{V_i} = \overline{\lambda_i} I_{V_i}.$$

We want to express T^* as a polynomial of T. Define

$$p(T) = \sum_{i=1}^{r} \overline{\lambda_i} \frac{(T-\lambda_1 I)\cdots(T-\lambda_{i-1}I)(T-\lambda_{i+1}I)\cdots(T-\lambda_r I)}{(\lambda_i-\lambda_1)\cdots(\lambda_i-\lambda_{i-1})(\lambda_i-\lambda_{i+1})\cdots(\lambda_i-\lambda_r)}.$$

For each $v_i \in E(\lambda_i, T)$,

$$\begin{aligned} p(T)v_i &= \overline{\lambda_i}\frac{(T-\lambda_1 I)\cdots(T-\lambda_{i-1}I)(T-\lambda_{i+1}I)\cdots(T-\lambda_r I)v_i}{(\lambda_i-\lambda_1)\cdots(\lambda_i-\lambda_{i-1})(\lambda_i-\lambda_{i+1})\cdots(\lambda_i-\lambda_r)} \\ &= \overline{\lambda_i}v_i = T^*v_i. \end{aligned}$$

T is normal implies T is diagonalisable. Pick a basis of eigenvectors $v_1,\ldots,v_n$ of T. Then $p(T)v_i = T^*v_i$ for $i = 1,\ldots,n$ implies $T^* = p(T)$. □

7B 10 11 12(use Q9 to prove)

Exercise 8.12 ([Axl24] 7B Q17)**.** Suppose $\mathbf{F} = \mathbb{R}$ and $\mathscr{E} \subset \mathscr{L}(V)$. Prove that there is an orthonormal basis of V with respect to which every element of $\mathscr{E}$ has a diagonal matrix if and only if S and T are commuting self-adjoint operators for all $S, T \in \mathscr{E}$.

This exercise extends the real spectral theorem to the context of a collection of commuting self-adjoint operators.

Exercise 8.13 ([Axl24] 7B Q19)**.** Suppose $T \in \mathscr{L}(V)$ is self-adjoint, and $U \leq V$ is invariant under T.

(i) Prove that $U^\perp$ is invariant under T.

(ii) Prove that $T|_U \in \mathscr{L}(V)$ is self-adjoint.

(iii) Prove that $T|_{U^\perp} \in \mathscr{L}(U^\perp)$ is self-adjoint.

Solution.

(i) Let $v \in U^\perp$. Then for all $w \in U$, $\langle v, w\rangle = 0$. Since U is invariant under V, $Tw \in U$, so

$$\langle Tv, w\rangle = \langle v, Tw\rangle = 0.$$

Thus $Tw \in U^\perp$. Hence $U^\perp$ is invariant under T.

(ii) For all $v, w \in U$, since $T \in \mathscr{L}(V)$ is self-adjoint,

$$\langle Tv, w\rangle = \langle v, Tw\rangle .$$

Restricting T to U gives

$$\langle T|_U v, w\rangle = \langle v, T|_U w\rangle .$$

Hence $T|_U$ is self-adjoint.

(iii) This follows from (i) and (ii).

□

Exercise 8.14 ([Axl24] 7B Q20)**.** Suppose $T \in \mathscr{L}(V)$ is normal, and $U \leq V$ is invariant under T.

(i) Prove that $U^\perp$ is invariant under T.

(ii) Prove that U is invariant under T^*.

(iii) Prove that $(T|_U)^* = (T^*)|_U$.

(iv) Prove that $T|_U \in \mathscr{L}(U)$ and $T|_{U^\perp} \in \mathscr{L}(U^\perp)$ are normal operators.

Solution.

(i) Let $v \in U^\perp$. Then $\langle v, w\rangle = 0$ for all $w \in U$. Since U is invariant under T, $Tw \in U$, so

$$\langle T^*v, w\rangle = \langle v, Tw\rangle = 0$$

which implies that $T^*v \in U^\perp$. Hence $U^\perp$ is invariant under T^*.

Using Exercise 9, T^* is a polynomial of T. Let $T = p(T^*)$, then $U^\perp$ is invariant under $p(T^*)$, which implies that $U^\perp$ is invariant under T.

Since T is normal, T is diagonalisable. Since U is invariant, the restriction $T|_U \in \mathscr{L}(U)$ is diagonalisable. Pick a basis of eigenvectors $\{u_1, \ldots, u_m\}$ of U. Then

$$Tu_1 = \lambda_1 u_1, \quad \ldots, \quad Tu_m = \lambda_m u_m.$$

T is normal implies

$$T^*u_1 = \overline{\lambda_1} u_1, \quad \ldots, \quad T^*u_m = \overline{\lambda_m} u_m.$$

For each $v \in U^\perp$, $\langle v, u_1\rangle = \cdots = \langle v, u_m\rangle = 0$, so

$$\langle Tv, u_i\rangle = \langle v, T^*u_i\rangle = \left\langle v, \overline{\lambda_i} u_i \right\rangle = \lambda_i \langle v, u_i\rangle = 0$$

for $i = 1, \ldots, n$. Thus $Tv \in U^\perp$. Hence $U^\perp$ is invariant under T.

(ii) This follows from (i).

(iii) We know $T|_U, T^*|_U \in \mathscr{L}(U)$. For all $v, w \in U$,

$$\langle Tv, w\rangle = \langle v, T^*w\rangle$$

so

$$\langle T|_U v, w\rangle = \langle v, T^*|_U w\rangle .$$

Hence $(T|_U)^* w = T^*|_U w$.

(iv) For each $v \in U$, $T^*Tv = TT^*v$ implies

$$T^*|_U T|_U v = T|_U T^*|_U v$$

so

$$(T|_U)^* T|_U v = T|_U (T|_U)^* v.$$

□

Exercise 8.15 ([Axl24] 7B Q21)**.** Suppose that T is a self-adjoint operator on a finite-dimensional inner product space, and that 2 and 3 are the only eigenvalues of T. Prove that

$$T^2 - 5T + 6I = 0.$$

We say a matrix A is *symmetric* if $A^\top = A$, and *Hermitian* if $A^* = A$.

Exercise 8.16 ([Axl24] 7B Q24)**.** Suppose U is a finite-dimensional vector space, and $T \in \mathscr{L}(U)$.

(i) Suppose $\mathbf{F} = \mathbb{R}$. Prove that T is diagonalisable if and only if there exists a basis of U such that the matrix of T with respect to this basis is symmetric.

(ii) Suppose $\mathbf{F} = \mathbb{C}$. Prove that T is diagonalisable if and only if there exists a basis of U such that the matrix of T with respect to this basis commutes with its conjugate transpose.

Solution.

(i)

(ii)

□

7C 1 3 5 6 7 11 13 14 15 16 17 18

Exercise 8.17 ([Axl24] 7C Q18)**.** Suppose S and T are positive operators on V. Prove that ST is a positive operator if and only if S and T commute.

Solution.

$\boxed{\Longrightarrow}$ Suppose ST is positive. Then ST is self-adjoint, so

$$ST = (ST)^* = T^*S^* = TS.$$

Thus S and T commute.

$\boxed{\Longleftarrow}$ Suppose $ST = TS$. Then

$$(ST)^* = T^*S^* = TS = ST$$

implies ST is self-adjoint. Let $v \in V$, then

$$\langle STv, v\rangle = \left\langle S\sqrt{T}v, \sqrt{T}v\right\rangle \geq 0,$$

since S is positive. □

Exercise 8.18 ([Axl24] 7C Q22)**.** Suppose $T \in \mathscr{L}(V)$ is a positive operator and $u \in V$ is such that $\|u\| = 1$, and $\|Tu\| \geq \|Tv\|$ for all $v \in V$ with $\|v\| = 1$. Show that u is an eigenvector corresponding to the largest eigenvalue of T.

Solution. Let $\{e_1, \ldots, e_n\}$ be the orthogonal eigenbasis of V, with corresponding values $\lambda_1, \ldots, \lambda_n$ sorted from smallest to largest.

Let $u = a_1e_1 + \cdots + a_ne_n$. Then

$$\begin{aligned} Tu &= a_1Te_1 + \cdots + a_nTe_n \\ &= a_1\lambda_1e_1 + \cdots + a_n\lambda_ne_n, \end{aligned}$$

so

$$\|Tu\|^2 = \left\|\sum_{i=1}^n a_i\lambda_ie_i\right\|^2 = \sum_{i=1}^n |a_i|^2{\lambda_i}^2.$$

We can take $v = e_n$, then we have that

$$\|Tu\|^2 = \sum_{i=1}^n |a_i|^2{\lambda_i}^2 \geq \sum_{i=1}^n |a_i|^2{\lambda_n}^2 = {\lambda_n}^2 = \|Tv\|^2$$

which shows the desired conclusion. □

Exercise 8.19 ([Axl24] 7C Q24)**.** Suppose $S, T \in \mathscr{L}(V)$ are positive operators. Prove that

$$\ker(S+T) = \ker S \cap \ker T.$$

Proof.

$\boxed{\supset}$

$$\begin{aligned} v \in \ker S \cap \ker T &\implies Sv = Tv = \mathbf{0} \\ &\implies (S+T)v = \mathbf{0} \\ &\implies v \in \ker(S+T) \end{aligned}$$

$\boxed{\subset}$

$$
\begin{aligned}
& v \in \ker(S+T) \\
\implies & (S+T)v = \mathbf{0} \\
\implies & 0 = \langle (S+T)v, v\rangle = \underbrace{\langle Sv, v\rangle}_{\geq 0} + \underbrace{\langle Tv, v\rangle}_{\geq 0} \\
\implies & \langle Sv, v\rangle = \langle Tv, v\rangle = 0 \\
\implies & \left\langle \sqrt{S}v, \sqrt{S}v \right\rangle = \left\langle \sqrt{T}v, \sqrt{T}v \right\rangle = 0 \\
\implies & \sqrt{S}v = \sqrt{T}v = \mathbf{0} \\
\implies & Sv = Tv = \mathbf{0}
\end{aligned}
$$

□

7D 2 3 5 9 11

Exercise 8.20 ([Axl24] 7D Q18)**.** Prove that if A is a symmetric matrix with real entries, then there exists a unitary matrix Q with real entries such that Q^*AQ is a diagonal matrix.

Solution. □

7E 1 2 4 7 8 9 10 11 13

7F 1-7 19

Exercise 8.21 ([Axl24] 7F Q8)**.**

(i) Prove that if $T \in \mathscr{L}(V)$ and $\|T\| < 1$, then $I+T$ is invertible.

(ii) Suppose that $S \in \mathscr{L}(V)$ is invertible. Prove that if $T \in \mathscr{L}(V)$ and $\|S-T\| < \dfrac{1}{\|S^{-1}\|}$, then T is invertible.

Solution.

(i) Let $v \in \ker(I+T)$. Then $(I+T)v = \mathbf{0}$, so $Tv = -v$. Taking the norm gives

$$\|v\| = \|Tv\| \leq \|T\|\|v\|.$$

Thus

$$(1 - \|T\|)\|v\| \leq 0.$$

Since $1 - \|T\| > 0$, we must have $\|v\| \leq 0$. Hence $v = \mathbf{0}$.

(ii) Let $v \in \ker T$. Then $Tv = \mathbf{0}$, so

$$Sv = (S-T)v.$$

Since S is invertible,

$$v = S^{-1}(S+T)v.$$

Taking norm,

$$\begin{aligned}\|v\| &= \left\|S^{-1}(S+T)v\right\| \\ &= \left\|S^{-1}\right\|\|S-T\|\|v\|.\end{aligned}$$

Thus

$$(\underbrace{1-\left\|S^{-1}\right\|\|S-T\|}_{>0})\|v\| \leq 0.$$

This implies $\|v\| = 0$, so $v = \mathbf{0}$. Hence T is injective, so T is invertible.

□

Then $B_{\frac{1}{\|S^{-1}\|}}(S)$ is an open ball in $\mathscr{L}(V)$ consisting of invertible linear maps. Hence the set of invertible operators in $\mathscr{L}(V)$ is an open subset of $\mathscr{L}(V)$.

Exercise 8.22 ([Axl24] 7F Q9)**.** Suppose $T \in \mathscr{L}(V)$. Prove that

$$\forall \varepsilon > 0, \quad \exists S \in \mathscr{L}(V) \text{ invertible}, \quad 0 < \|T-S\| < \varepsilon.$$

Solution. Define $S = T + \delta I$ for some $0 < \delta < \varepsilon$. Then we have

$$\|T-S\| = \|\delta I\| = \delta$$

which satisifes the desired condition. Note that if T is invertible, we can simply choose a sufficiently small $\delta < 1/\left\|T^{-1}\right\|$; if not, then any $\delta \in (0,1)$ can make S invertible. □

Exercise 8.23 ([Axl24] 7F Q10)**.** Suppose $\dim V > 1$ and $T \in \mathscr{L}(V)$ is not invertible. Prove that

$$\forall \varepsilon > 0, \quad \exists S \in \mathscr{L}(V) \text{ not invertible}, \quad 0 < \|T-S\| < \varepsilon.$$

Solution. □

Exercise 8.24 ([Axl24] 7F Q11)**.** Suppose $\mathbf{F} = \mathbb{C}$ and $T \in \mathscr{L}(V)$. Prove that

$$\forall \varepsilon > 0, \quad \exists S \in \mathscr{L}(V) \text{ diagonalisable}, \quad 0 < \|T-S\| < \varepsilon.$$

Exercise 8.25 ([Axl24] 7F Q12)**.** Suppose $T \in \mathscr{L}(V)$ is a positive operator. Show that $\left\|\sqrt{T}\right\| = \sqrt{\|T\|}$.

Solution. Let $\|T\| = \sigma_1$, the largest singular value of T. Then $\left\|\sqrt{T}\right\| = \sqrt{\sigma_1} = \sqrt{\|T\|}$. □

Exercise 8.26 ([Axl24] 7F Q14)**.** Suppose $U, W \leq V$ are such that $\|P_U - P_W\| < 1$. Prove that $\dim U = \dim W$.

Solution. We prove the contrapositive. Suppose $\dim U < \dim W$. Then the orthogonal complement

$$\dim U^\perp = \dim V - \dim U > \dim V - \dim W,$$

or

$$\dim W + \dim U^\perp > \dim W.$$

We know

$$\begin{aligned}\dim V &\geq \dim(W + U^\perp) \\ &= \dim W + \dim U^\perp - \dim(W \cap U^\perp) \\ &> \dim V - \dim(W \cap U^\perp)\end{aligned}$$

which implies $\dim W \cap U^\perp > 0$. Pick $v \in W \cap U^\perp$, $\|v\| = 1$. Then $P_U v = \mathbf{0}$ and $P_W v = v$, so

$$\|(P_U - P_w)v\| = \|v\| = 1.$$

Hence $\|P_U - P_W\| \geq 1$, so $\dim U \geq \dim W$. Similarly, $\|P_W - P_U\| \leq 1$, so $\dim W \geq \dim U$. Therefore $\dim U = \dim W$. □

Exercise 8.27 ([Axl24] 7F Q19)**.** Prove that if $T \in \mathscr{L}(V, W)$, then $\|T^*T\| = \|T\|^2$.

Solution. Let ${s_1}^2$ be the greatest eigenvalue of T^*T. Then $\|T\| = s_1$, so $\|T\|^2 = {s_1}^2 = \|T^*T\|$. □

9 Operators on Complex Vector Spaces

9.1 Generalised Eigenvectors and Nilpotent Operators

Kernels of Powers of an Operator

We begin this chapter with a study of kernels of powers of an operator. The following result provides a sequence of increasing kernels.

Lemma 9.1. *Suppose $T \in \mathscr{L}(V)$. Then*

$$\{\mathbf{0}\} = \ker T^0 \subset \ker T^1 \subset \cdots \subset \ker T^k \subset \ker T^{k+1} \subset \cdots.$$

Proof. Let $v \in \ker T^k$, for non-negative integer k. Then $T^k v = \mathbf{0}$, so $T^{k+1} v = T\left(T^k v\right) = T(\mathbf{0}) = \mathbf{0}$. Thus $v \in \ker T^{k+1}$.

Hence $\ker T^k \subset \ker T^{k+1}$ for non-negative integers k. □

The next result states that if two consecutive terms in the sequence are equal, then all later terms are equal.

Lemma 9.2. *Suppose $T \in \mathscr{L}(V)$, and $\ker T^m = \ker T^{m+1}$ for some non-negative integer m. Then*

$$\ker T^m = \ker T^{m+1} = \ker T^{m+1} = \ker T^{m+3} = \cdots.$$

Proof. Let k be a positive integer. We want to prove that

$$\ker T^{m+k} = \ker T^{m+k+1}.$$

$\boxed{\subset}$ This follows from 9.1.

$\boxed{\supseteq}$ Let $v \in \ker T^{m+k+1}$. Then

$$T^{m+1}(T^k v) = T^{m+k+1} v = \mathbf{0}.$$

Hence

$$T^k v \in \ker T^{m+1} = \ker T^m.$$

Thus $T^{m+k}v = T^m(T^k v) = \mathbf{0}$, which means that $v \in \ker T^{m+k}$. □

The result above raises the question of whether there exists a non-negative integer m such that $\ker T^m = \ker T^{m+1}$. The next result shows that this equality holds, at least when $m = \dim V$.

Lemma 9.3 (Kernels stop growing). *Suppose $T \in \mathscr{L}(V)$. Then*

$$\ker T^{\dim V} = \ker T^{\dim V+1} = \ker T^{\dim V+2} = \cdots.$$

Proof. By 9.2, it suffices to prove that $\ker T^{\dim V} = \ker T^{\dim V+1}$. Suppose, for a contradiction, that $\ker T^{\dim V} \neq \ker T^{\dim V+1}$. Then

$$\{\mathbf{0}\} = \ker T^0 \subsetneq \ker T^1 \subsetneq \cdots \subsetneq \ker T^{\dim V} \subsetneq \ker T^{\dim V+1},$$

where we have strict inclusions in the chain above. At each of the strict inclusions, the dimension increases by at least 1. Thus $\dim \ker T^{\dim V+1} \geq \dim V + 1$. Since $\ker T^{\dim V+1}$ is a subspace of V, it cannot have a larger dimension than the whole space. Hence we have reached a contradiction. □

It is not true that $V = \ker T \oplus \operatorname{im} T$ for every $T \in \mathscr{L}(V)$. However, the next result can be a useful substitute.

Proposition 9.4 (Direct sum decomposition). *Suppose $T \in \mathscr{L}(V)$. Then*

$$V = \ker T^{\dim V} \oplus \operatorname{im} T^{\dim V}.$$

Proof. Let $n = \dim V$.

- We first show that $\ker T^n + \operatorname{im} T^n$ is a direct sum. By 3.13, we will show that

  $$\ker T^n \cap \operatorname{im} T^n = \{\mathbf{0}\}.$$

  Let $v \in \ker T^n \cap \operatorname{im} T^n$. Then $T^n v = \mathbf{0}$, and there exists $u \in V$ such that $v = T^n u$. Applying T^n to both sides gives

  $$T^n v = T^{2n} u = \mathbf{0}$$

  which implies

  $$T^n u = \mathbf{0}.$$

Thus $v = T^n u = \mathbf{0}$.

- Next we show that $V = \ker T^n \oplus \operatorname{im} T^n$. We shall use 3.33 to show that the subspace $\ker T^n \oplus \operatorname{im} T^n$ equals the whole space V:

$$\begin{aligned} \dim(\ker T^n \oplus \operatorname{im} T^n) &= \dim \ker T^n + \dim \operatorname{im} T^n && \text{[by 4.50]} \\ &= \dim V && \text{[by fundamental theorem of linear maps]} \end{aligned}$$

□

Generalised Eigenvectors

Definition 9.5 (Generalised eigenvector)**.** Suppose $T \in \mathscr{L}(V)$, and $\lambda \in \mathbf{F}$ is an eigenvalue of T. We say $v \in V \setminus \{\mathbf{0}\}$ is a ***generalised eigenvector*** of T corresponding to λ, if

$$(T - \lambda I)^k v = \mathbf{0}$$

for some positive integer k.

That is, there exists $k \in \mathbb{Z}^+$ such that

$$v \in \ker(T - \lambda I)^k.$$

Remark. If $k = 1$, then this coincides with the usual eigenvector. Hence all eigenvectors are generalised eigenvectors.

Remark. We do not define generalised eigenvalues because they are no different from the usual eigenvalues. Reason: if $(T - \lambda I)^k$ is not injective for some positive integer k, then $T - \lambda I$ is not injective, and hence λ is an eigenvalue of T.

A non-zero vector $v \in V$ is a generalised eigenvector of T corresponding to λ if and only if

$$(T - \lambda I)^{\dim V} v = \mathbf{0},$$

as follows from applying 9.1 and 9.3 to the operator $T - \lambda I$.

As we know, an operator on a complex vector space may not have enough eigenvectors to form a basis of the domain. The next result shows that on a complex vector space, there are enough generalised eigenvectors to do this.

Proposition 9.6. *Suppose $\mathbf{F} = \mathbb{C}$, and $T \in \mathscr{L}(V)$. Then T has a basis of generalised eigenvectors in V.*

Proof. Let $n = \dim V$. We shall induct on n.

If $n = 1$, then every non-zero vector in V is an eigenvector of T. Thus the desired result holds.

Suppose $n > 1$, and the desired result holds for all smaller values of $\dim V$. Let λ be an eigenvalue of T. Applying 9.4 to $T - \lambda I$ shows that

$$V = \ker(T - \lambda I)^n \oplus \operatorname{im}(T - \lambda I)^n.$$

If $\ker(T - \lambda I)^n = V$, then every non-zero vector in V is a generalised eigenvector of T, and thus in this case there is a basis of V consisting of generalised eigenvectors of T. Hence we can assume that $\ker(T - \lambda I)^n \neq V$, which implies that $\operatorname{im}(T - \lambda I)^n \neq \{\mathbf{0}\}$. Since λ is an eigenvalue of T, we have $\ker(T - \lambda I)^n \neq \{\mathbf{0}\}$. Thus

$$0 < \dim \operatorname{im}(T - \lambda I)^n < n.$$

Note that $\operatorname{im}(T - \lambda I)^n$ is invariant under T. Let $S \in \mathscr{L}(\operatorname{im}(T - \lambda I)^n)$ be defined by

$$S = T|_{\operatorname{im}(T - \lambda I)^n}.$$

By induction hypothesis applied to S, there is a basis of $\operatorname{im}(T - \lambda I)^n$ consisting of generalised eigenvectors of S, which of course are generalised eigenvectors of T. Adjoining that basis of $\operatorname{im}(T - \lambda I)^n$ to a basis of $\ker(T - \lambda I)^n$ gives a basis of V consisting of generalised eigenvectors of T. □

Suppose $T \in \mathscr{L}(V)$. If v is an eigenvector of T, then the corresponding eigenvalue λ is uniquely determined by the equation $Tv = \lambda v$, which can be satisfied by only one $\lambda \in \mathbf{F}$ (since $v \neq \mathbf{0}$). The next result shows a similar result holds for generalised eigenvectors: if v is a generalised eigenvector of T, then the equation $(T - \lambda I)^{\dim V} v = \mathbf{0}$ can be satisfied by only one $\lambda \in \mathbf{F}$.

Lemma 9.7. *Suppose $T \in \mathscr{L}(V)$. Then each generalised eigenvector of T corresponds to only one eigenvalue of T.*

Proof. Suppose $T \in \mathscr{L}(V)$. Let $v \in V$ be a generalised eigenvector of T corresponding to eigenvalues λ and λ'.

Let m be the smallest positive integer such that $(T - \lambda' I)^m v = \mathbf{0}$. Let $n = \dim V$. Then

$$\begin{aligned}
\mathbf{0} &= (T - \lambda I)^n v \\
&= \big((T - \lambda' I) + (\lambda' - \lambda) I\big)^n v \\
&= \sum_{i=1}^{n} \binom{n}{i} (\lambda' - \lambda)^{n-i} (T - \lambda' I)^i v.
\end{aligned}$$

Applying the operator $(T - \lambda I)^{m-1}$ to both sides gives

$$\mathbf{0} = (\lambda' - \lambda)^n (T - \lambda' I)^{m-1} v.$$

Since $(T-\lambda' I)^{m-1}v \neq \mathbf{0}$, the equation above implies that $\lambda' = \lambda$, as desired. □

Recall that by 6.5, eigenvectors corresponding to distinct eigenvalues are linearly independent. We now prove a similar result for generalised eigenvectors.

Proposition 9.8. *Suppose $T \in \mathscr{L}(V)$. Then every set of generalised eigenvectors of T corresponding to distinct eigenvalues of T is linearly independent.*

Proof. Suppose, for a contradiction, that the desired result is false. Then there exists a smallest positive integer m such that $\{v_1,\dots,v_m\}$ are linearly dependent generalised eigenvectors of T corresponding to distinct eigenvalues $\lambda_1,\dots,\lambda_m$ of T.

Thus there exist $a_1,\dots,a_m \in \mathbf{F}$, none of which are 0 (because of the minimality of m), such that

$$a_1v_1 + \cdots + a_mv_m = \mathbf{0}.$$

Let $n = \dim V$. Applying $(T-\lambda_m I)^n$ to both sides gives

$$a_1(T-\lambda_m I)^n v_1 + \cdots + a_{m-1}(T-\lambda_m I)^n v_{m-1} = \mathbf{0}. \tag{I}$$

Let $i \in \{1,\dots,m-1\}$. Then

$$(T-\lambda_m I)^n v_i \neq \mathbf{0},$$

because otherwise v_i would be a generalised eigenvector of T corresponding to distinct eigenvalues λ_i and λ_m, contradicting 8.11. However

$$(T-\lambda_i I)^n\left((T-\lambda_m I)^n v_i\right) = (T-\lambda_m I)^n\left((T-\lambda_i I)^n v_i\right) = \mathbf{0}.$$

Thus the last two equations show that $(T-\lambda_m I)^n v_i$ is a generalised eigenvector of T corresponding to the eigenvalue λ_i. Hence

$$(T-\lambda_m I)^n v_1,\dots,(T-\lambda_m I)^n v_{m-1}$$

is a linearly dependent set (by (I)) of $m-1$ generalised eigenvectors corresponding to distinct eigenvalues, contradicting the minimality of m. □

Nilpotent Operators

Definition 9.9 (Nilpotent operator)**.** An operator is ***nilpotent*** if some power of it equals 0.

Thus an operator $T \in \mathscr{L}(V)$ is nilpotent if and only if every non-zero vector in V is a generalised eigenvector of T corresponding to the eigenvalue 0.

Lemma 9.10. *Suppose $T \in \mathscr{L}(V)$ is nilpotent. Then $T^{\dim V} = 0$.*

Proof. Since T is nilpotent, there exists $k \in \mathbb{Z}^+$ such that $T^k = 0$.

Thus $\ker T^k = V$. Now 9.1 and 9.3 imply that $\ker T^{\dim V} = V$. Hence $T^{\dim V} = 0$. □

The next result concerns the eigenvalues of nilpotent operators.

Proposition 9.11. *Suppose $T \in \mathscr{L}(V)$.*

(i) If T is nilpotent, then 0 is the only eigenvalue of T.

(ii) If $\mathbf{F} = \mathbb{C}$ and 0 is the only eigenvalue of T, then T is nilpotent.

Proof.

(i) Suppose T is nilpotent. Then $T^m = 0$ for some $m \in \mathbb{Z}^+$. This implies that T is not injective. Thus 0 is an eigenvalue of T.

To show that T has no other eigenvalues, suppose λ is an eigenvalue of T. Then there exists $v \in V \setminus \{\mathbf{0}\}$ such that

$$\lambda v = Tv.$$

Repeatedly applying T to both sides of the equation gives

$$\lambda^m v = T^m v = \mathbf{0}.$$

Thus $\lambda = 0$, as desired.

(ii) Suppose $\mathbf{F} = \mathbb{C}$, and 0 is the only eigenvalue of T. By 6.14, the minimal polynomial of T equals z^m for some positive integer m. Thus $T^m = 0$. Hence T is nilpotent.

□

The next result provides a characterisation of a nilpotent operator, in terms of its minimal polynomial and matrix.

Proposition 9.12. *Suppose $T \in \mathscr{L}(V)$. Then the following are equivalent:*

(i) T is nilpotent.

(ii) The minimal polynomial of T is z^m, for some positive integer m.

(iii) T has a strictly upper-triangular matrix with respect to some basis of V.

Proof.

$\boxed{\text{(i)} \Longrightarrow \text{(ii)}}$ Suppose T is nilpotent. Then $T^n = 0$ for some $n \in \mathbb{Z}^+$.

By 6.15, z^n is a polynomial multiple of the minimal polynomial of T. Thus the minimal polynomial of T is z^m for some $m \in \mathbb{Z}^+$.

$\boxed{\text{(ii)} \Longrightarrow \text{(iii)}}$ Suppose the minimal polynomial of T must be z^m, for some $m \in \mathbb{Z}^+$.

Since 0 is the only zero of z^m, by 6.14, 0 is the only eigenvalue of T.

Thus by 6.24, T has an upper-triangular matrix with respect to some basis of V.

By 6.22, since the eigenvalues of T are the diagonal entries of the matrix, we conclude that all entries on the diagonal of this matrix are 0.

$\boxed{\text{(iii)} \Longrightarrow \text{(i)}}$ Suppose T has a strictly upper-triangular matrix with respect to some basis of V.

Then the diagonal entries are all 0. By 6.22, $T^{\dim V} = 0$. Hence T is nilpotent. □

9.2 Generalised Eigenspace Decomposition

Generalised Eigenspaces

Definition 9.13 (Generalised eigenspace)**.** Suppose $T \in \mathscr{L}(V)$, and $\lambda \in \mathbf{F}$. The ***generalised eigenspace*** of T corresponding to λ is

$$G(\lambda, T) := \left\{ v \in V \mid \exists k \in \mathbb{Z}^+, (T - \lambda I)^k v = \mathbf{0} \right\}.$$

That is, the generalised eigenspace of T corresponding to λ is the set of generalised eigenvectors of T corresponding to λ, along with $\mathbf{0}$.

Remark. $E(\lambda, T) \subset G(\lambda, T)$.

A consequence of the next result is $G(\lambda, T)$ is a subspace of V.

Lemma 9.14. *Suppose $T \in \mathscr{L}(V)$, and $\lambda \in \mathbf{F}$. Then*

$$G(\lambda, T) = \ker(T - \lambda I)^{\dim V}.$$

Proof.

$\boxed{\supset}$ Let $v \in \ker(T - \lambda I)^{\dim V}$. Then $(T - \lambda I)^{\dim V} = \mathbf{0}$. By definition, $v \in G(\lambda, T)$.

$\boxed{\subset}$ Let $v \in G(\lambda, T)$. Then $(T - \lambda I)^k v = \mathbf{0}$ for some $k \in \mathbb{Z}^+$, so $v \in \ker(T - \lambda I)^k$.

From 9.1 and 9.3 (with $T - \lambda I$ replacing T), we get $v \in \ker(T - \lambda I)^{\dim V}$. □

Lemma 9.15. *Suppose $\mathbf{F} = \mathbb{C}$ and $T \in \mathscr{L}(V)$. Let $\lambda_1, \ldots, \lambda_m$ be distinct eigenvalues of T. Then*

(i) $G(\lambda_i, T)$ is invariant under T, for each $i = 1, \ldots, m$.

(ii) $(T - \lambda_i I)|_{G(\lambda_i, T)}$ is nilpotent, for each $i = 1, \ldots, m$.

Proof.

(i) Let $i \in \{1, \ldots, m\}$. Then

$$G(\lambda_i, T) = \ker(T - \lambda_i I)^{\dim V}$$

which is invariant under T.

(ii) Let $i \in \{1, \ldots, m\}$. Let $v \in G(\lambda_i, T)$, then $(T - \lambda_i I)^{\dim V} v = \mathbf{0}$. Thus

$$\left((T - \lambda_i I)|_{G(\lambda_i, T)}\right)^{\dim V} = \mathbf{0}.$$

Hence $(T-\lambda_i I)|_{G(\lambda_i,T)}$ is nilpotent.

□

Theorem 9.16 (Generalised eigenspace decomposition). *Suppose* $\mathbf{F}=\mathbb{C}$ *and* $T\in\mathscr{L}(V)$. *Let* $\lambda_1,\ldots,\lambda_m$ *be distinct eigenvalues of* T. *Then*

$$V=G(\lambda_1,T)\oplus\cdots\oplus G(\lambda_m,T). \tag{9.1}$$

Proof. We first show $G(\lambda_1,T)+\cdots+G(\lambda_m,T)$ is a direct sum. Suppose

$$v_1+\cdots+v_m=\mathbf{0}$$

where each $v_i\in G(\lambda_i,T)$. Since generalised eigenvectors of T corresponding to distinct eigenvalues are linearly independent (by 9.8), this implies that each $v_i=\mathbf{0}$.

By 9.6, each vector in V can be written as a finite sum of generalised eigenvectors of T. Hence

$$V=G(\lambda_1,T)\oplus\cdots\oplus G(\lambda_m,T).$$

□

Multiplicity of an Eigenvalue

Definition 9.17 (Multiplicity). Suppose $T\in\mathscr{L}(V)$. The (algebraic) ***multiplicity*** of an eigenvalue λ of T is defined as

$$\dim G(\lambda,T).$$

Equivalently, the multiplicity of an eigenvalue λ of T equals

$$\dim\ker(T-\lambda I)^{\dim V}.$$

The next result states that the sum of the multiplicities of all eigenvalues of T equals $\dim V$.

Lemma 9.18. *Suppose* $\mathbf{F}=\mathbb{C}$ *and* $T\in\mathscr{L}(V)$. *Let* $\lambda_1,\ldots,\lambda_m$ *be the distinct eigenvalues of* T, *with multiplicities* $d_1,\ldots,d_m$. *Then*

$$d_1+\cdots+d_m=\dim V.$$

Proof. By the generalised eigenspace decomposition,

$$V=G(\lambda_1,T)\oplus\cdots\oplus G(\lambda_m,T).$$

Then the dimension of the direct sum is

$$\begin{aligned}\dim V &= \dim G(\lambda_1, T) \oplus \cdots \oplus G(\lambda_m, T) \\ &= \dim G(\lambda_1, T) + \cdots + \dim G(\lambda_m, T) \\ &= d_1 + \cdots + d_m.\end{aligned}$$

□

Definition 9.19 (Characteristic polynomial)**.** Suppose $\mathbf{F} = \mathbb{C}$ and $T \in \mathscr{L}(V)$. The ***characteristic polynomial*** of T is

$$q(z) = (z - \lambda_1)^{d_1} \cdots (z - \lambda_m)^{d_m}$$

where $\lambda_1, \ldots, \lambda_m$ are the distinct eigenvalues of T, with multiplicities $d_1, \ldots, d_m$.

Note that the characteristic polynomial of T has degree $d_1 + \cdots + d_m = \dim V$; the zeros of the characteristic polynomial are the eigenvalues of T, by definition.

Remark. The condition that $\mathbf{F} = \mathbb{C}$ is required; for general fields $\mathbf{F}$, we define the characteristic polynomial as

$$q(z) = \det(\lambda I - T).$$

Theorem 9.20 (Cayley–Hamilton theorem)**.** *Suppose $\mathbf{F} = \mathbb{C}$ and $T \in \mathscr{L}(V)$. Let q be the characteristic polynomial of T. Then*

$$q(T) = 0.$$

Proof. Let $\lambda_1, \ldots, \lambda_m$ be the distinct eigenvalues of T, with multiplicities $d_1, \ldots, d_m$. For each $i \in \{1, \ldots, m\}$, we know that $(T - \lambda_i I)|_{G(\lambda_i, T)}$ is nilpotent. Thus

$$(T - \lambda_i I)^{d_i}|_{G(\lambda_i, T)} = 0,$$

since $d_i := \dim G(\lambda_i, T)$.

By the generalised eigenspace decomposition, to prove that $q(T) = 0$, it suffices to show that $q(T)|_{G(\lambda_i, T)} = 0$ for each i. Fix $i \in \{1, \ldots, m\}$. We have

$$q(T) = (T - \lambda_1 I)^{d_1} \cdots (T - \lambda_m I)^{d_m}.$$

Since all the operators on the RHS all commute, we can move the factor $(T - \lambda_i I)^{d_i}$ to be the last term. Since $(T - \lambda_i I)^{d_i}|_{G(\lambda_i, T)} = 0$, we have $q(T)|_{G(\lambda_i, T)} = 0$, as desired. □

An immediate corollary that the characteristic polynomial is a multiple of minimal polynomial. Thus if the minimal polynomial has degree $\dim V$, then it equals the characteristic polynomial.

Theorem 9.21. *Suppose $\mathbf{F} = \mathbb{C}$ and $T \in \mathscr{L}(V)$. Suppose $\mathscr{M}(T)$ is upper-triangular with respect to some basis of V. Then the number of times that each eigenvalue λ of T appears on the diagonal of $\mathscr{M}(T)$ equals its multiplicity.*

Proof. Suppose $\{v_1, \dots, v_n\}$ is a basis of V, with respect to which $\mathscr{M}(T)$ is upper-triangular; denote $A = \mathscr{M}(T)$. Let $\lambda_1, \dots, \lambda_n$ denote the entries on the diagonal of A. Thus for each $i = 1, \dots, n$,

$$Tv_i = u_i + \lambda_i v_i$$

for some $u_i \in \operatorname{span}(v_1, \dots, v_{i-1})$. Hence if $\lambda_i \neq 0$, then Tv_i is not a linear combination of $Tv_1, \dots, Tv_{i-1}$. By the linear dependence lemma, the set of those Tv_i such that $\lambda_i \neq 0$ is linearly independent.

Let d denote the number of indices $i \in \{1, \dots, n\}$ such that $\lambda_i = 0$. Then the conclusion of the previous paragraph implies that

$$\dim \operatorname{im} T \geq n - d.$$

By the fundamental theorem of linear maps, the inequality above implies that

$$\dim \ker T \leq d. \tag{I}$$

The matrix of the operator T^n with respect to the basis $\{v_1, \dots, v_n\}$ is the upper-triangular matrix A^n, which has diagonal entries $\lambda_1^n, \dots, \lambda_n^n$ [see Exercise 2(b) in Section 5C]. Since $\lambda_i^n = 0$ if and only if $\lambda_i = 0$, the number of times that 0 appears on the diagonal of A^n equals d. Thus applying (I) with T replaced with T^n, we have

$$\dim \ker T^n \leq d. \tag{II}$$

Let m_λ denote the multiplicity of an eigenvalue λ of T; let d_λ denote the number of times that λ appears on the diagonal of A.

Claim. $m_\lambda = d_\lambda$, for each eigenvalue λ of T.

Replacing T in (II) with $T - \lambda I$, we see that

$$m_\lambda \leq d_\lambda$$

for each eigenvalue λ of T. Summing both sides over all eigenvalues λ of T,

$$n = \sum_\lambda m_\lambda \leq \sum_\lambda d_\lambda = n$$

where the first equality holds since the sum of multiplicities of eigenvalues equals $n = \dim V$; the second equality holds since the the diagonal of A has length n.

Hence equality in $m_\lambda \leq d_\lambda$ holds for each eigenvalue λ of T, as desired. □

Block Diagonal Matrices

Often we can understand a matrix better by thinking of it as composed of smaller matrices.

Definition 9.22 (Block diagonal matrix). A ***block diagonal matrix*** is a square matrix of the form

$$\begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_m \end{pmatrix}$$

where $A_1, \ldots, A_m$ are square matrices lying along the diagonal, and all other entries of the matrix equal 0.

Proposition 9.23. *Suppose $\mathbf{F} = \mathbb{C}$ and $T \in \mathscr{L}(V)$. Let $\lambda_1, \ldots, \lambda_m$ be distinct eigenvalues of T, with multiplicities $d_1, \ldots, d_m$. Then there is a basis of V with respect to which T has a block diagonal matrix of the form*

$$\begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_m \end{pmatrix}$$

where each A_i is a $d_i \times d_i$ upper-triangular matrix of the form

$$A_i = \begin{pmatrix} \lambda_i & & * \\ & \ddots & \\ 0 & & \lambda_i \end{pmatrix}.$$

Proof. Each $(T - \lambda_i I)|_{G(\lambda_i, T)}$ is nilpotent. For each i, choose a basis of $G(\lambda_i, T)$, which is a vector space of dimension d_i, such that the matrix of $(T - \lambda_i I)|_{G(\lambda_i, T)}$ with respect to this basis is as in 8.18(c).

Thus with respect to this basis, the matrix of $T|_{G(\lambda_i, T)}$, which equals $(T - \lambda_i I)|_{G(\lambda_i, T)} + \lambda_i I|_{G(\lambda_i, T)}$, looks like the desired form shown above for A_i.

The generalised eigenspace decomposition shows that putting together the bases of the $G(\lambda_i, T)$'s chosen above gives a basis of V. The matrix of T with respect to this basis has the desired form. □

9.3 Consequences of Generalised Eigenspace Decomposition

Square Roots of Operators

We begin by showing that the identity plus any nilpotent operator has a square root.

Lemma 9.24. *Suppose $T \in \mathscr{L}(V)$ is nilpotent. Then $I+T$ has a square root.*

Proof. Our motivation is the Taylor series for $\sqrt{1+x}$:

$$\sqrt{1+x} = 1 + a_1 x + a_2 x^2 + \cdots \tag{I}$$

for some coefficients $a_1, a_2, \ldots$.

Since T is nilpotent, $T^m = 0$ for some $m \in \mathbb{Z}^+$. In (I), suppose we replace x with T, and 1 with I. Then the infinite sum on the RHS becomes a finite sum, since $T^k = 0$ for all $k \geq m$. Thus we guess that there is a square root of $I+T$ of the form

$$1 + a_1 T + a_2 T^2 + \cdots + a_{m-1} T^{m-1}.$$

Having made this guess, we can try to choose $a_1, a_2, \ldots, a_{m-1}$ such that the operator above has its square equal to $I+T$. Now

$$\begin{aligned}(1 + a_1 T + a_2 T^2 + a_3 T^3 + \cdots + a_{m-1} T^{m-1})^2 \\ = I + 2a_1 T + (2a_2 + a_1^2) T^2 + (2a_3 + 2a_1 a_2) T^3 \\ + (2a_{m-1} + \text{terms involving } a_1, \ldots, a_{m-2}) T^{m-1}.\end{aligned}$$

We want the RHS of the equation above to equal $I+T$. Hence choose a_1 such that $2a_1 = 1$ (thus $a_1 = 1/2$). Next, choose a_2 such that $2a_2 + a_1^2 = 0$ (thus $a_2 = -1/8$). Then choose a_3 such that the coefficient of T^3 on the RHS of the equation above equals 0 (thus $a_3 = 1/16$).

Continue in this fashion for each $i = 4, \ldots, m-1$, at each step solving for a_i so that the coefficient of T^i on the RHS of the equation above equals 0. Actually we do not care about the explicit formula for the a_i's. We only need to know that some choice of the a_i's gives a square root of $I+T$. □

Proposition 9.25. *Suppose V is a complex vector space and $T \in \mathscr{L}(V)$ is invertible. Then T has a square root.*

Remark. This result holds only on complex vector spaces. For example, the operator of multiplication by -1 on the one-dimensional real vector space $\mathbb{R}$ has no square root.

Proof. Let $\lambda_1, \ldots, \lambda_m$ be the distinct eigenvalues of T. For each $i = 1, \ldots, m$, there exists a nilpotent operator $T_i \in \mathscr{L}(G(\lambda_i, T))$ such that $T|_{G(\lambda_i,T)} = \lambda_i I + T_i$. Since T is invertible, none of the λ_i's equals 0, so we can write

$$T|_{G(\lambda_i,T)} = \lambda_i \left(I + \frac{T_i}{\lambda_i} \right).$$

Since T_i/λ_i is nilpotent, $I + T_i/\lambda_i$ has a square root (by the lemma). Thus $T|_{G(\lambda_i,T)}$ has a square root

$$R_i = \sqrt{\lambda_i} \sqrt{I + \frac{T_i}{\lambda_i}}.$$

By the generalised eigenspace decomposition, every $v \in V$ can be written uniquely in the form

$$v = u_1 + \cdots + u_m,$$

where each $u_i \in G(\lambda_i, T)$. Using this decomposition,

Claim. Define $R \in \mathscr{L}(V)$ by

$$Rv = R_1 u_1 + \cdots + R_m u_m.$$

Then R is a square root of T.

□

By imitating the techniques in this section, you should be able to prove that if V is a complex vector space and $T \in \mathscr{L}(V)$ is invertible, then T has a k-th root for every positive integer k.

Jordan Normal Form

By 9.23, we know that if V is a complex vector space, then every $T \in \mathscr{L}(V)$ has a nice upper-triangular matrix with respect to some basis of V.

In this subsection, we will see that we can do even better – there is a basis of V with respect to which the matrix of T contains 0's everywhere except possibly on the diagonal and the line directly above the diagonal.

Definition 9.26 (Jordan basis). Suppose $T \in \mathscr{L}(V)$. We say a basis of V is a ***Jordan basis*** for T if with respect to this basis T has a block diagonal matrix

$$\begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_p \end{pmatrix}$$

in which each A_i is an upper-triangular matrix of the form

$$A_i = \begin{pmatrix} \lambda_i & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_i \end{pmatrix}.$$

Most of the work in proving that every operator on a finite-dimensional complex vector space has a Jordan basis occurs in proving the special case below of nilpotent operators.

Proposition 9.27. *Every nilpotent operator has a Jordan basis.*

Proof. Induct on $\dim V$.

If $\dim V = 1$, then the only nilpotent operator is the 0 operator, so the desired result holds.

Suppose $\dim V > 1$ and the desired result holds on all vector spaces of smaller dimension.

Since T is nilpotent, let m be the smallest positive integer such that $T^m = 0$. Thus there exists $u \in V$ such that $T^{m-1}u \neq \mathbf{0}$. Let

$$U = \operatorname{span}\left(u, Tu, \ldots, T^{m-1}u\right).$$

Note that $\{u, Tu, \ldots, T^{m-1}u\}$ is linearly independent (why?). If $U = V$, then writing this set in reverse order gives a Jordan basis for T, and we are done (why?). Thus assume $U \neq V$.

Note that U is invariant under T. By induction hypothesis, there is a basis of U that is a Jordan basis for $T|_U$.

Idea. We will find a subspace W of V such that W is also invariant under T, and $V = U \oplus W$. Again by induction hypothesis, there is a basis of W that is a Jordan basis for $T|_W$. Putting together the Jordan bases for $T|_U$ and $T|_W$ gives a Jordan basis for T.

Let $\phi \in V'$ be such that $\phi\left(T^{m-1}u\right) \neq 0$.

Claim. $W = \left\{v \in V \mid \phi(T^k v) = 0 \quad (k = 0, \ldots, m-1)\right\}$.

- If $v \in W$, then $\phi\left(T^k(Tv)\right) = 0$ for $k = 0, \ldots, m-1$ (the case $k = m-1$ holds since $T^m = 0$). Thus $Tv \in W$. Hence W is a subspace of V that is invariant under T.
- To show $U + W$ is a direct sum, let $v \in U \cap W$ with $v \neq \mathbf{0}$. Since $v \in U$, there exist $a_0, \ldots, a_{m-1} \in \mathbf{F}$ such that

$$v = a_0 u + a_1 Tu + \cdots + a_{m-1} T^{m-1} u.$$

Let $i \in \{0,\dots,m-1\}$ be the smallest index such that $a_i \neq 0$. Applying T^{m-i-1} to both sides gives

$$T^{m-i-1}v = a_i T^{m-1}u$$

where we have used $T^m = 0$. Now apply ϕ to both sides to get

$$\phi\left(T^{m-i-1}v\right) = a_i\phi\left(T^{m-1}u\right) \neq 0.$$

This implies $v \notin W$. Hence $U \cap W = \{\mathbf{0}\}$, so $U + W$ is a direct sum.

- To show $V = U \oplus W$, define $S\colon V \to \mathbf{F}^m$ by

$$Sv = \left(\phi(v), \phi(Tv), \dots, \phi(T^{m-1}v)\right).$$

Thus $\ker S = W$. Hence

$$\dim W = \dim\ker S = \dim V - \dim\operatorname{im} S \geq \dim V - m.$$

This implies

$$\dim(U \oplus W) = \dim U + \dim W \geq m + (\dim V - m) = \dim V,$$

so $\dim(U + W) = \dim V$. Hence $V = U \oplus W$.

□

Theorem 9.28 (Jordan form). *Suppose* $\mathbf{F} = \mathbb{C}$. *Then every operator has a Jordan basis.*

Proof. Let $\lambda_1,\dots,\lambda_m$ be the distinct eigenvalues of T. By the generalised eigenspace decomposition,

$$V = G(\lambda_1, T) \oplus \cdots \oplus G(\lambda_m, T)$$

and each $(T - \lambda_i I)|_{G(\lambda_i,T)}$ is nilpotent. By 9.27, some basis of each $G(\lambda_i, T)$ is a Jordan basis for $(T - \lambda_i I)|_{G(\lambda_i,T)}$. Put these bases together to get a basis of V that is a Jordan basis of T. □

9.4 Trace

Definition 9.29 (Trace). Suppose $A \in \mathscr{M}_{n\times n}(\mathbf{F})$. The ***trace*** of A is the sum of the diagonal entries of A:

$$\operatorname{tr}(A) := \sum_{i=1}^{n} A_{ii}.$$

It is easy to verify that

$$\begin{aligned}\operatorname{tr}(A+B) &= \operatorname{tr}A + \operatorname{tr}B,\\ \operatorname{tr}(kA) &= k\operatorname{tr}A.\end{aligned}$$

Hence $\operatorname{tr}\colon \mathscr{M}_{n\times n}(\mathbf{F}) \to \mathbf{F}$ is linear.

Matrix multiplication is not commutative, but the next result shows that the order of matrix multiplication does not matter to the trace.

Lemma 9.30. *Suppose $A \in \mathscr{M}_{m\times n}(\mathbf{F})$, $B \in \mathscr{M}_{n\times m}(\mathbf{F})$. Then*

$$\operatorname{tr}(AB) = \operatorname{tr}(BA).$$

Proof. We have

$$\begin{aligned}\operatorname{tr}(AB) &= \sum_{i=1}^{n} (AB)_{ii}\\ &= \sum_{i=1}^{n} \left(\sum_{k=1}^{n} A_{ik}B_{ki} \right)\\ &= \sum_{k=1}^{n} \sum_{i=1}^{n} B_{ki}A_{ik}\\ &= \sum_{k=1}^{n} (BA)_{kk} = \operatorname{tr}(BA).\end{aligned}$$

□

The next result states that the trace of the matrix of an operator does not depend on the choice of basis.

Lemma 9.31. *Suppose $T \in \mathscr{L}(V)$. Suppose $\{u_1,\dots,u_n\}$ and $\{v_1,\dots,v_n\}$ are bases of V. Then*

$$\operatorname{tr}\mathscr{M}(T;\{u_1,\dots,u_n\}) = \operatorname{tr}\mathscr{M}(T;\{v_1,\dots,v_n\}).$$

Proof. Let $A = \mathscr{M}(T;\{u_1,\dots,u_n\})$ and $B = \mathscr{M}(T;\{v_1,\dots,v_n\})$. By the change-of-basis formula, there exists an invertible $n \times n$ matrix C such that $A = C^{-1}BC$. Thus

$$\begin{aligned}\operatorname{tr} A &= \operatorname{tr}\big((C^{-1}B)C\big)\\ &= \operatorname{tr}\big(C(C^{-1}B)\big)\\ &= \operatorname{tr}\big((CC^{-1})B\big)\\ &= \operatorname{tr} B.\end{aligned}$$

□

The previous result allows us to define the trace of an operator to be the trace of its matrix with respect to any basis.

Definition 9.32 (Trace). Suppose $T \in \mathscr{L}(V)$. The ***trace*** of T is

$$\operatorname{tr} T := \operatorname{tr} \mathscr{M}(T;\{v_1,\dots,v_n\})$$

where $\{v_1,\dots,v_n\}$ is any basis of V.

It is easy to verify that for all $T,S \in \mathscr{L}(V)$, $k \in \mathbf{F}$,

$$\begin{aligned}\operatorname{tr}(T+S) &= \operatorname{tr} T + \operatorname{tr} S,\\ \operatorname{tr}(kT) &= k \operatorname{tr} T.\end{aligned}$$

Hence $\operatorname{tr}\colon \mathscr{L}(V) \to \mathbf{F}$ is linear. In addition $\operatorname{tr}(TS) = \operatorname{tr}(ST)$.

Lemma 9.33. *Suppose $\mathbf{F} = \mathbb{C}$ and $T \in \mathscr{L}(V)$. Then $\operatorname{tr} T$ equals the sum of the eigenvalues of T, with each eigenvalue included as many times as its multiplicity.*

Proof. By 9.23, there exista a basis of V, with respect to which T has an upper-triangular matrix, where the diagonal entries are the eigenvalues of T, with each eigenvalue included as many times as its multiplicity.

The result follows from the definition of the trace of an operator. □

Lemma 9.34. *Suppose $\mathbf{F} = \mathbb{C}$ and $T \in \mathscr{L}(V)$. Let $n = \dim V$. Then $\operatorname{tr} T$ equals the negative of the coefficient of z^{n-1} in the characteristic polynomial of T.*

Proof. Suppose $\lambda_1,\dots,\lambda_n$ are the eigenvalues of T, with each eigenvalue included as many times as its multiplicity. Then the characteristic polynomial of T equals

$$(z-\lambda_1)\cdots(z-\lambda_n).$$

On expanding, the coefficient of z^{n-1} is $-(\lambda_1 + \cdots + \lambda_n)$. But $\operatorname{tr} T = \lambda_1 + \cdots + \lambda_n$, so we are done. □

The next result gives a formula for the trace of an operator on an inner product space.

Lemma 9.35. *Suppose V is an inner product space, $T \in \mathscr{L}(V)$, and $\{e_1, \ldots, e_n\}$ is an orthonormal basis of V. Then*

$$\operatorname{tr} T = \sum_{i=1}^{n} \langle Te_i, e_i \rangle .$$

Proof. Consider $\mathscr{M}(T; \{e_1, \ldots, e_n\})$. For $i = 1, \ldots, n$,

$$Te_i = \mathscr{M}(T)_{1,i} e_1 + \cdots + \mathscr{M}(T)_{n,i} e_n.$$

Thus

$$\langle Te_i, e_i \rangle = \mathscr{M}(T)_{i,i}.$$

Hence the desired follows. □

Lemma 9.36. *The function $\operatorname{tr} \colon \mathscr{L}(V) \to \mathbf{F}$ is a linear functional on $\mathscr{L}(V)$, such that*

$$\operatorname{tr}(ST) = \operatorname{tr}(TS)$$

for all $S, T \in \mathscr{L}(V)$.

Proof. Choose a basis of V. All matrices of operators in this proof will be with respect to that basis. Suppose $S, T \in \mathscr{L}(V)$.

(i) If $\lambda \in \mathbf{F}$, then $\operatorname{tr}(\lambda T) = \operatorname{tr} \mathscr{M}(\lambda T) = \operatorname{tr}(\lambda \mathscr{M}(T)) = \lambda \operatorname{tr} \mathscr{M}(T) = \lambda \operatorname{tr} T$.

(ii) We have $\operatorname{tr}(S+T) = \operatorname{tr} \mathscr{M}(S+T) = \operatorname{tr}(\mathscr{M}(S) + \mathscr{T}) = \operatorname{tr} \mathscr{M}(S) + \operatorname{tr} \mathscr{M}(T) = \operatorname{tr} S + \operatorname{tr} T$.

In addition, $\operatorname{tr}(ST) = \operatorname{tr} \mathscr{M}(ST) = \operatorname{tr}(\mathscr{M}(S)\mathscr{M}(T)) = \operatorname{tr}(\mathscr{T}\mathscr{S}) = \operatorname{tr} \mathscr{M}(TS) = \operatorname{tr}(TS)$. □

Corollary 9.37. *Suppose V is finite-dimensional. There do not exist operators $S, T \in \mathscr{L}(V)$ such that $ST - TS = I$.*

Proof. Suppose $S, T \in \mathscr{L}(V)$. Then

$$\operatorname{tr}(ST - TS) = \operatorname{tr}(ST) - \operatorname{tr}(TS) = 0.$$

The trace of I equals $\dim V$, which is not 0. Since $ST - TS$ and I have different traces, they cannot be equal. □

Exercises

Exercise 9.1 ([Axl24] 8A Q2)**.** Suppose $T \in \mathscr{L}(V)$, m is a positive integer, $v \in V$, and $T^{m-1}v \neq \mathbf{0}$ but $T^m v = \mathbf{0}$. Prove that $v, Tv, T^2v, \ldots, T^{m-1}v$ is linearly independent.

Exercise 9.2 ([Axl24] 8A Q4)**.** Suppose $T \in \mathscr{L}(V)$, $\lambda \in \mathbf{F}$, and m is a positive integer such that the minimal polynomial of T is a polynomial multiple of $(z-\lambda)^m$. Prove that

$$\dim \ker (T - \lambda I)^m \geq m.$$

Exercise 9.3 ([Axl24] 8A Q5)**.** Suppose $T \in \mathscr{L}(V)$ and m is a positive integer. Prove that

$$\dim \ker T^m \leq m \dim \ker T.$$

Exercise 9.4 ([Axl24] 8A Q6)**.** Suppose $T \in \mathscr{L}(V)$. Show that

$$V = \operatorname{im} T^0 \supset \operatorname{im} T^1 \supset \cdots \supset \operatorname{im} T^k \supset \operatorname{im} T^{k+1} \supset \cdots.$$

Exercise 9.5 ([Axl24] 8A Q7)**.** Suppose $T \in \mathscr{L}(V)$ and m is a non-negative integer such that

$$\operatorname{im} T^m = \operatorname{im} T^{m+1}.$$

Prove that $\operatorname{im} T^k = \operatorname{im} T^m$ for all $k > m$.

Exercise 9.6 ([Axl24] 8A Q8)**.** Suppose $T \in \mathscr{L}(V)$. Prove that

$$\operatorname{im} T^{\dim V} = \operatorname{im} T^{\dim V + 1} = \operatorname{im} T^{\dim V + 2} = \cdots.$$

Exercise 9.7 ([Axl24] 8A Q9)**.** Suppose $T \in \mathscr{L}(V)$ and m is a non-negative integer. Prove that

$$\ker T^m = \ker T^{m+1} \iff \operatorname{im} T^m = \operatorname{im} T^{m+1}.$$

Exercise 9.8 ([Axl24] 8A Q11)**.** Suppose $T \in \mathscr{L}(V)$. Prove that there is a basis of V consisting of generalised eigenvectors of T if and only if the minimal polynomial of T equals $(z-\lambda_1)\cdots(z-\lambda_m)$ for some $\lambda_1, \ldots, \lambda_m \in \mathbf{F}$.

Exercise 9.9 ([Axl24] 8A Q12)**.** Suppose $T \in \mathscr{L}(V)$ is such that every vector in V is a generalised eigenvector of T. Prove that there exists $\lambda \in \mathbf{F}$ such that $T - \lambda I$ is nilpotent.

Exercise 9.10 ([Axl24] 8A Q13)**.** Suppose $S, T \in \mathscr{L}(V)$. Prove that if ST is nilpotent, then TS is nilpotent.

Solution. Suppose ST is nilpotent. Then $(ST)^m = 0$ for some $m \in \mathbb{Z}^+$. Thus

$$(TS)^{m+1} = T(ST)^m S = 0.$$

Hence TS is nilpotent. □

Exercise 9.11 ([Axl24] 8A Q14)**.** Suppose $T \in \mathscr{L}(V)$ is nilpotent and $T \neq 0$. Prove that T is not diagonalisable.

Exercise 9.12 ([Axl24] 8A Q15)**.** Suppose $\mathbf{F} = \mathbb{C}$ and $T \in \mathscr{L}(V)$. Prove that T is diagonalisable if and only if every generalised eigenvector of T is an eigenvector of T.

Exercise 9.13 ([Axl24] 8A Q16)**.**

(i) Give an example of nilpotent operators S, T on the same vector space such that neither $S+T$ nor ST is nilpotent.

(ii) Suppose $S, T \in \mathscr{L}(V)$ are nilpotent and $ST = TS$. Prove that $S+T$ and ST are nilpotent.

Solution.

(i)

(ii) Let $n = \dim V$. Then $S^n = T^n = 0$. We have

$$(ST)^n = S^n T^n = 0,$$
$$(S+T)^{2n} = \sum_{i=0}^{2n} \binom{2n}{i} S^i T^{2n-i} = 0.$$

□

Exercise 9.14 ([Axl24] 8A Q17)**.** Suppose $T \in \mathscr{L}(V)$ is nilpotent and m is a positive integer such that $T^m = 0$. Prove that $I - T$ is invertible and

$$(1-T)^{-1} = I + T + \cdots + T^{m-1}.$$

Exercise 9.15 ([Axl24] 8A Q18)**.** Suppose $T \in \mathscr{L}(V)$ is nilpotent. Prove that $T^{1+\dim \operatorname{im} T} = 0$.

Solution. Let $n = \dim V$. Consider the tower

$$\{\mathbf{0}\} \subsetneq \ker T \subsetneq \ker T^2 \subsetneq \cdots.$$

Then $0 < \dim\ker T < \dim\ker T^2 < \cdots$, so

$$\begin{aligned}\dim\ker T^{1+k} &\geq \dim\ker T^k + 1\\ &\geq \dim\ker T^{k-1} + 2\\ &\geq \cdots \geq \dim\ker T + k.\end{aligned}$$

Take $k = \dim\operatorname{im} T$. Then

$$\dim\ker T^{1+\dim\operatorname{im} T} \geq \dim\ker T + \dim\operatorname{im} T = \dim V.$$

But $\ker T^{1+\dim\operatorname{im} T}$ is a subspace of V, so $\ker T^{1+\dim\operatorname{im} T} = V$. Hence $T^{1+\dim\operatorname{im} T} = 0$. □

Exercise 9.16 ([Axl24] 8A Q19)**.** Suppose $T \in \mathscr{L}(V)$ is not nilpotent. Show that

$$V = \ker T^{\dim V - 1} \oplus \operatorname{im} T^{\dim V - 1}.$$

Solution. Let $n = \dim V$. Consider the tower

$$\{\mathbf{0}\} \subsetneq \ker T \subsetneq \cdots \subsetneq \ker T^m = \ker T^{m+1} = \cdots = \ker T^n = \cdots.$$

Then $\dim\ker T^m \geq m$. Since T is not nilpotent, $\dim\ker T^m < n$. Thus $m < n$, so $m \leq n-1$.

We now show the direct sum. Let $v \in \ker T^{n-1} \cap \operatorname{im} T^{n-1}$. Then $T^{n-1}v = \mathbf{0}$ and $v = T^{n-1}u$ for some $u \in V$. We have

$$T^{2(n-1)}u = \mathbf{0} \implies u \in \ker T^{2n-1} = \ker T^{n-1}$$

which implies $v = T^{n-1}u = \mathbf{0}$. Hence

$$V = \ker T^{\dim V - 1} \oplus \operatorname{im} T^{\dim V - 1}.$$

□

Exercise 9.17 ([Axl24] 8A Q20)**.** Suppose V is an inner product space and $T \in \mathscr{L}(V)$ is normal and nilpotent. Prove that $T = 0$.

Solution. Let $n = \dim V$. Since T is normal, $T^*T = TT^*$, so

$$(T^*T)^n = (T^*)^n T^n = 0$$

since T is nilpotent.

Let $S = T^*T$, then

$$\begin{aligned} S \text{ is self-adjoint} &\implies S \text{ is diagonalisable} \\ &\implies S = 0 && [\because S \text{ is nilpotent}] \\ &\implies T^*T = 0 \\ &\implies T = 0 \end{aligned}$$

where the reason for the last line is as follows: for all $v \in V$, $\langle Tv, Tv\rangle = \langle v, T^*Tv\rangle = 0 \implies \|Tv\| = 0 \implies Tv = \mathbf{0}$. □

Exercise 9.18 ([Axl24] 8A Q21)**.** Suppose $T \in \mathscr{L}(V)$ is such that $\ker T^{\dim V - 1} \neq \ker T^{\dim V}$. Prove that T is nilpotent and that $\dim \ker T^k = k$ for every integer k with $0 \leq k \leq \dim V$.

Solution. Consider the tower

$$0 \subsetneq \ker T \subsetneq \cdots \subsetneq \ker T^m = \ker T^{m+1} = \cdots$$

for some $m \in \mathbb{Z}^+$. Since $\ker T^{\dim V - 1} \neq \ker T^{\dim V}$, this implies $m \geq \dim V$. But

$$0 < \dim \ker T < \cdots < \dim \ker T^m.$$

Thus $\dim \ker T^m \geq m \geq \dim V$. Since $\ker T^m$ is a subspace of V, we must have $\dim \ker T^m = \dim V$, so $V = \ker T^m$. Thus equality holds, which implies $m = \dim V$.

Note that

$$0 < \dim \ker T < \cdots < \dim \ker T^{\dim V} = \dim V$$

is a strictly increasing sequence of positive integers of length $\dim V$, we must have $\dim \ker T^k = k$ for $k = 1, \ldots, n$ (by the pigeonhole principle). □

Exercise 9.19 ([Axl24] 8A Q22)**.** Suppose $T \in \mathscr{L}(\mathbb{C}^5)$ is such that $\operatorname{im} T^4 \neq \operatorname{im} T^5$. Prove that T is nilpotent.

Solution. Use the previous exercise. □

Exercise 9.20 ([Axl24] 8A Q23)**.** Give an example of an operator T on a finite-dimensional real vector space such that 0 is the only eigenvalue of T but T is not nilpotent.

Exercise 9.21 ([Axl24] 8B Q2)**.** Suppose $T \in \mathscr{L}(V)$ is invertible. Prove that $G(\lambda, T) = G\left(\frac{1}{\lambda}, T^{-1}\right)$ for every $\lambda \in \mathbf{F}$ with $\lambda \neq 0$.

Exercise 9.22 ([Axl24] 8B Q3)**.** Suppose $T \in \mathscr{L}(V)$. Suppose $S \in \mathscr{L}(V)$ is invertible. Prove that T and $S^{-1}TS$ have the same eigenvalues with the same multiplicities.

Exercise 9.23 ([Axl24] 8B Q4)**.** Suppose $\dim V \geq 2$ and $T \in \mathscr{L}(V)$ is such that $\ker T^{\dim V - 2} \neq \ker T^{\dim V - 1}$. Prove that T has at most two distinct eigenvalues.

Exercise 9.24 ([Axl24] 8B Q5)**.** Suppose $T \in \mathscr{L}(V)$ and 3 and 8 are eigenvalues of T. Let $n = \dim V$. Prove that $V = \ker T^{n-2} \oplus \operatorname{im} T^{n-2}$.

Exercise 9.25 ([Axl24] 8B Q6)**.** Suppose $T \in \mathscr{L}(V)$ and λ is an eigenvalue of T. Explain why the exponent of $z - \lambda$ in the factorisation of the minimal polynomial of T is the smallest positive integer m such that $(T - \lambda I)^m|_{G(\lambda,T)} = 0$.

Exercise 9.26 ([Axl24] 8B Q7)**.** Suppose $T \in \mathscr{L}(V)$ and λ is an eigenvalue of T with multiplicity d. Prove that $G(\lambda, T) = \ker(T - \lambda I)^d$.

Exercise 9.27 ([Axl24] 8B Q8)**.** Suppose $T \in \mathscr{L}(V)$ and $\lambda_1, \ldots, \lambda_m$ are the distinct eigenvalues of T. Prove that

$$V = G(\lambda_1, T) \oplus \cdots \oplus G(\lambda_m, T)$$

if and only if the minimal polynomial of T equals $(z - \lambda_1)^{k_1} \cdots (z - \lambda_m)^{k_m}$ for some positive integers $k_1, \ldots, k_m$.

Exercise 9.28 ([Axl24] 8B Q9)**.** Suppose $\mathbf{F} = \mathbb{C}$ and $T \in \mathscr{L}(V)$. Prove that there exist $D, N \in \mathscr{L}(V)$ such that $T = D + N$, the operator D is diagonalisable, N is nilpotent, and $DN = ND$.

Exercise 9.29 ([Axl24] 8B Q10)**.** Suppose V is a complex inner product space, $\{e_1, \ldots, e_n\}$ is an orthonormal basis of V, and $T \in \mathscr{L}(V)$. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of T, each included as many times as its multiplicity. Prove that

$$|\lambda_1|^2 + \cdots + |\lambda_n|^2 \leq \|Te_1\|^2 + \cdots + \|Te_n\|^2.$$

Solution. By the singular value decomposition, we have

$$\begin{aligned}\sum_{i=1}^{n} \|Te_i\|^2 &= \sum_{i=1}^{n} \|U\Sigma V^* e_i\|^2 \\ &= \sum_{i=1}^{n} \|U\Sigma f_i\|^2 \\ &= \sum_{i=1}^{n} \|\Sigma f_i\|^2 \\ &= \sum_{i=1}^{n} \|\lambda_i f_i\|^2 \\ &\geq \sum_{i=1}^{n} |\lambda_i|^2\end{aligned}$$

where the last line follows from Bessel's inequality. □

Exercise 9.30 ([Axl24] 8B Q11)**.** Give an example of an operator on $\mathbb{C}^4$ whose characteristic polynomial equals $(z-7)^2(z-8)^2$.

Exercise 9.31 ([Axl24] 8B Q12)**.** Give an example of an operator on $\mathbb{C}^4$ whose characteristic polynomial equals $(z-1)(z-5)^3$ and whose minimal polynomial equals $(z-1)(z-5)^2$.

Exercise 9.32 ([Axl24] 8B Q13)**.** Give an example of an operator on $\mathbb{C}^4$ whose characteristic and minimal polynomials both equal $z(z-1)^2(z-3)$.

Exercise 9.33 ([Axl24] 8B Q14)**.** Give an example of an operator on $\mathbb{C}^4$ whose characteristic polynomial equals $z(z-1)^2(z-3)$ and whose minimal polynomial equals $z(z-1)(z-3)$.

Exercise 9.34 ([Axl24] 8B Q17)**.** Suppose $\mathbf{F} = \mathbb{C}$ and $P \in \mathscr{L}(V)$ is such that $P^2 = P$. Prove that the characteristic polynomial of P is $z^m(z-1)^n$, where $m = \dim \ker P$ and $n = \dim \operatorname{im} P$.

Solution. Since $P^2 = P$, we have

$$V = \ker P \oplus \operatorname{im} P.$$

Note that $P|_{\ker P} = 0$.

Let $v \in \operatorname{im} P$. Then $v = Pw$ for some $w \in V$. Thus $Pv = P^2 w = Pw = v$.

Hence $\operatorname{im} P = E(1,P) = G(0,P)$ and $\ker P = E(0,P) = G(1,P)$, by the direct sum decomposition above.

Therefore the characteristic polynomial is

$$\begin{aligned} q(z) &= (z-0)^{\dim G(0,P)}(z-1)^{\dim G(1,P)} \\ &= z^{\dim \ker P}(z-1)^{\dim \operatorname{im} P} \\ &= z^m(z-1)^n. \end{aligned}$$

□

Exercise 9.35 ([Axl24] 8B Q18)**.** Suppose $T \in \mathscr{L}(V)$ and λ is an eigenvalue of T. Explain why the following four numbers equal each other:

(i) The exponent of $z-\lambda$ in the factorisation of the minimal polynomial of T.

(ii) The smallest positive integer m such that $(T-\lambda I)^m|_{G(\lambda,T)} = 0$.

(iii) The smallest positive integer m such that

$$\ker(T-\lambda I)^m = \ker(T-\lambda I)^{m+1}.$$

(iv) The smallest positive integer m such that

$$\operatorname{im}(T-\lambda I)^m = \operatorname{im}(T-\lambda I)^{m+1}.$$

Exercise 9.36 ([Axl24] 8B Q20)**.** Suppose that $\mathbf{F} = \mathbb{C}$ and $V_1, \ldots, V_m$ are non-zero subspaces of V such that

$$V = V_1 \oplus \cdots \oplus V_m.$$

Suppose $T \in \mathscr{L}(V)$ and each V_i is invariant under T. For each i, let p_i denote the characteristic polynomial of $T|_{V_i}$. Prove that the characteristic polynomial of T equals $p_1 \cdots p_m$.

Exercise 9.37 ([Axl24] 8B Q21)**.** Suppose $p, q \in \mathbb{C}[z]$ are monic polynomials with the same zeros, and q is a polynomial multiple of p. Prove that there exists $T \in \mathscr{L}(\mathbb{C}^{\deg q})$ such that the characteristic polynomial of T is q, and the minimal polynomial of T is p.

Exercise 9.38 ([Axl24] 8B Q22)**.** Suppose A and B are block diagonal matrices of the form

$$A = \begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_m \end{pmatrix}, \quad B = \begin{pmatrix} B_1 & & 0 \\ & \ddots & \\ 0 & & B_m \end{pmatrix},$$

where A_i and B_i are square matrices of the same size for each $i = 1, \ldots, m$. Show that AB

is a block diagonal matrix of the form

$$AB = \begin{pmatrix} A_1B_1 & & 0 \\ & \ddots & \\ 0 & & A_mB_m \end{pmatrix}.$$

Exercise 9.39 ([Axl24] 8C Q1)**.**

8C 1-14

Exercise 9.40 ([Axl24] 8D Q1)**.** Suppose V is an inner product space and $v, w \in V$. Define $T \in \mathscr{L}(V)$ by $Tu = \langle u, v\rangle w$. Find a formula for $\operatorname{tr} T$.

Solution. Let $\{e_1, \ldots, e_n\}$ be an orthonormal basis of V. Then

$$\begin{aligned} \operatorname{tr} T &= \sum_{i=1}^{n} \langle Te_i, e_i\rangle \\ &= \sum_{i=1}^{n} \langle \langle e_i, v\rangle w, e_i\rangle \\ &= \sum_{i=1}^{n} \langle e_i, v\rangle \langle w, e_i\rangle \\ &= \sum_{i=1}^{n} v_i w_i \\ &= v \cdot w. \end{aligned}$$

□

Exercise 9.41 ([Axl24] 8D Q2)**.** Suppose $P \in \mathscr{L}(V)$ satisfies $P^2 = P$. Prove that

$$\operatorname{tr} P = \dim \operatorname{im} P.$$

Solution. Note that $\operatorname{tr} P = \sum_{i=1}^{n} \lambda_i$ where $\lambda_i = 1$ or 0. The multiplicity of $\lambda_i = 1$ determines the $\dim \operatorname{im} P$ and thus gives the desired conclusion. □

Exercise 9.42 ([Axl24] 8D Q4)**.** Suppose V is an inner product space and $T \in \mathscr{L}(V)$. Prove that

$$\operatorname{tr} T^* = \overline{\operatorname{tr} T}.$$

Solution. Let $\{e_1, \ldots, e_n\}$ be an orthonormal basis of V. Then

$$\begin{aligned}\operatorname{tr} T^* &= \sum_{i=1}^{n} \langle T^* e_i, e_i \rangle = \sum_{i=1}^{n} \langle e_i, T e_i \rangle \\ &= \sum_{i=1}^{n} \overline{\langle T e_i, e_i \rangle} = \overline{\sum_{i=1}^{n} \langle T e_i, e_i \rangle} = \overline{\operatorname{tr} T}.\end{aligned}$$

□

Exercise 9.43 ([Axl24] 8D Q5)**.** Suppose V is an inner product space. Suppose $T \in \mathscr{L}(V)$ is a positive operator and $\operatorname{tr} T = 0$. Prove that $T = 0$.

Solution. Since T is a positive operator, $\langle Tv, v \rangle \geq 0$ for all $v \in V$. Pick an orthonormal basis $\{e_1, \ldots, e_n\}$ of V. We have

$$\operatorname{tr} T = \sum_{i=1}^{n} \langle T e_i, e_i \rangle = 0.$$

But each $\langle T e_i, e_i \rangle \geq 0$. Thus we must have $\langle T e_i, e_i \rangle = 0$ for all $i = 1, \ldots, n$. This implies $\left\langle \sqrt{T} e_i, \sqrt{T} e_i \right\rangle = 0$, so $\sqrt{T} e_i = 0$ for all $i = 1, \ldots, n$. Hence $\sqrt{T} = 0$, so $T = 0$. □

Exercise 9.44 ([Axl24] 8D Q6)**.** Suppose V is an inner product space and $P, Q \in \mathscr{L}(V)$ are orthogonal projections. Prove that $\operatorname{tr}(PQ) \geq 0$.

Solution. Let $\{e_1, \ldots, e_n\}$ be an orthonormal basis of V. Then

$$\operatorname{tr}(PQ) = \sum_{i=1}^{n} \langle PQ e_i, e_i \rangle .$$

Since $\ker Q$ and $\operatorname{im} Q$ are orthogonal, we have the direct sum

$$V = \ker Q \oplus \operatorname{im} Q.$$

Let $\{e_1, \ldots, e_r\}$ be an orthonormal basis of $\operatorname{im} Q$, and $\{e_{r+1}, \ldots, e_n\}$ be an orthonormal basis of $\ker Q$. Then $\{e_1, \ldots, e_n\}$ is an orthonormal basis of V. Therefore

$$\begin{aligned}\operatorname{tr}(PQ) &= \sum_{i=1}^{n} \langle PQ e_i, e_i \rangle \\ &= \sum_{i=1}^{r} \langle P e_i, e_i \rangle \\ &= \sum_{i=1}^{r} \left\langle P^2 e_i, e_i \right\rangle \\ &= \sum_{i=1}^{r} \| P e_i \|^2 \geq 0.\end{aligned}$$

□

Exercise 9.45 ([Axl24] 8D Q8)**.** Prove or disprove: If $S,T \in \mathscr{L}(V)$, then $\operatorname{tr}(ST) = (\operatorname{tr} S)(\operatorname{tr} T)$.

Idea. Take $\operatorname{tr} S = 0$ but $\operatorname{tr} ST = 0$.

Solution. Suppose the matrices of S and T are $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$. Then ST has matrix $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

Then $(\operatorname{tr} S)(\operatorname{tr} T) = 0$ but $\operatorname{tr} ST = 2$. □

Exercise 9.46 ([Axl24] 8D Q9)**.** Suppose $T \in \mathscr{L}(V)$ is such that $\operatorname{tr}(ST) = 0$ for all $S \in \mathscr{L}(V)$. Prove that $T = 0$.

Solution. Let $\{e_1, \dots, e_n\}$ be an orthonormal basis of V.

Define S_{ij} to be such that maps e_j to e_i while keep all other zero.

Therefore, $\operatorname{tr}\left(S_{ij}T\right) = T_{ij} = 0$ for all i, j. Hence, we proved $T = 0$. □

Exercise 9.47 ([Axl24] 8D Q10)**.** Prove that the trace is the only linear functional $\tau \colon \mathscr{L}(V) \to \mathbf{F}$ such that

$$\tau(ST) = \tau(TS)$$

for all $S,T \in \mathscr{L}(V)$ and $\tau(I) = \dim V$.

Exercise 9.48 ([Axl24] 8D Q11)**.** Suppose V and W are inner product spaces and $T \in \mathscr{L}(V,W)$. Prove that if $\{e_1, \dots, e_n\}$ is an orthonormal basis of V, and $\{f_1, \dots, f_m\}$ is an orthonormal basis of W, then

$$\operatorname{tr}(T^*T) = \sum_{i=1}^{n} \sum_{j=1}^{m} |\langle Te_i, f_j \rangle|^2.$$

Solution. We have

$$\begin{aligned} \operatorname{tr}(T^*T) &= \sum_{i=1}^{n} \langle T^*Te_i, e_i \rangle \\ &= \sum_{i=1}^{n} \langle Te_i, Te_i \rangle = \sum_{i=1}^{n} \|Te_i\|^2 \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} |\langle Te_i, f_j \rangle|^2 \end{aligned}$$

where the last line follows from Pythagoras' theorem. □

Exercise 9.49 ([Axl24] 8D Q12, Frobenius norm)**.** Suppose V and W are finite-dimensional vector spaces.

(i) Prove that $\langle S,T\rangle = \operatorname{tr}(T^*S)$ defines an inner product on $\mathscr{L}(V,W)$.

(ii) Suppose $\{e_1,\dots,e_n\}$ is an orthonormal basis of V, and $\{f_1,\dots,f_m\}$ is an orthonormal basis of W.

Show that the inner product on $\mathscr{L}(V,W)$ from (i) is the same as the standard inner product on $\mathbf{F}^{mn}$, where we identify each element of $\mathscr{L}(V,W)$ with its matrix and then with an element of $\mathbf{F}^{mn}$.

Solution.

(i)
- $\langle T,T\rangle = \operatorname{tr}(T^*T) = \sum_{i=1}^{n}\sum_{j=1}^{m} |\langle Te_i, f_j\rangle|^2 \geq 0$, where equality holds $\iff \langle Te_i,f_j\rangle = 0 (\forall i,j) \iff T = 0$. Hence positive definitiveness holds.
- $\langle S_1+S_2,T\rangle = \operatorname{tr}(T^*(S_1+S_2)) = \operatorname{tr}(T^*S_1+T^*S_2) = \operatorname{tr}(T^*S_1)+\operatorname{tr}(T^*S_2) = \langle S_1,T\rangle + \langle S_2,T\rangle$.
  $\langle \lambda S,T\rangle = \operatorname{tr}(T^*(\lambda S)) = \operatorname{tr}(\lambda T^*S) = \lambda \operatorname{tr}(T^*S) = \lambda \langle S,T\rangle$. Hence sesquilinearity holds.
- $\langle T,S\rangle = \operatorname{tr}(S^*T) = \operatorname{tr}((T^*S)^*) = \overline{\operatorname{tr}(T^*S)} = \langle S,T\rangle$. Hence conjugate symmetry holds.

(ii) The standard inner product on $\mathbf{F}^{mn}$ for the two matrices A and B is

$$\langle A,B\rangle = \sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}\overline{B_{ij}}$$

which is exactly how we define in (i).

□

10 Multilinear Algebra and Determinants

10.1 Bilinear Forms and Quadratic Forms

Bilinear Forms

So far, we have been looking at *linear* things only. This can get quite boring. For a change, we look at *bilinear* maps instead.

Definition 10.1 (Bilinear form)**.** A ***bilinear form*** on V is a function $\beta : V \times V \to \mathbf{F}$ which is linear in each slot:

$$\begin{aligned} \beta(\cdot, u) &: V \to \mathbf{F}, \quad v \mapsto \beta(v, u) \\ \beta(u, \cdot) &: V \to \mathbf{F}, \quad v \mapsto \beta(u, v) \end{aligned}$$

are linear, for every $u \in V$.

Example.

- If V is a real inner product space, then the inner product $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ is a bilinear form.

  However, if V is a complex inner product space, then the inner product $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{C}$ is not a bilinear form, due to conjugate linearity in the second slot.

- Define $\beta : \mathbf{F}^3 \times \mathbf{F}^3 \to \mathbf{F}$ by

  $$\beta\left((x_1, x_2, x_3), (y_1, y_2, y_3)\right) = x_1 y_2 - 5 x_2 y_3 + 2 x_3 y_1.$$

  Then β is a bilinear form on $\mathbf{F}^3$.

Proof. We have

$$\begin{aligned}\beta(ax+x',y) &= \beta((ax_1+x'_1,ax_2+x'_2,ax_3+x'_3),(y_1,y_2,y_3))\\ &= (ax_1+x'_1)y_2-5(ax_2+x'_2)y_3+2(ax_3+x'_3)y_1\\ &= a(x_1y_2-5x_2y_3+2x_3y_1)+(x'_1y_2-5x'_2y_3+2x'_3y_1)\\ &= a\beta(x,y)+\beta(x',y)\end{aligned}$$

and

$$\begin{aligned}\beta(x,ay+y') &= \beta((x_1,x_2,x_3),(ay_1+y'_1,ay_2+y'_2,ay_3+y'_3))\\ &= x_1(ay_2+y'_2)-5x_2(ay_3+y'_3)+2x_3(ay_1+y'_1)\\ &= a(x_1y_2-5x_2y_3+2x_3y_1)+(x_1y'_2-5x_2y'_3+2x_3y'_1)\\ &= a\beta(x,y)+\beta(x,y').\end{aligned}$$

□

- Suppose $A \in \mathscr{M}_{n\times n}(\mathbf{F})$. Define $\beta_A : \mathbf{F}^n \times \mathbf{F}^n \to \mathbf{F}$ by

$$\beta_A((x_1,\ldots,x_n),(y_1,\ldots,y_n)) = \sum_{j=1}^{n}\sum_{i=1}^{n} A_{i,j}x_iy_j.$$

Then β_A is a bilinear form on $\mathbf{F}^n$.

Proof. We have

$$\begin{aligned}\beta_A(ax+x',y) &= \sum_{j=1}^{n}\sum_{i=1}^{n} A_{i,j}(ax_i+x'_i)y_j\\ &= \sum_{j=1}^{n}\sum_{i=1}^{n} A_{i,j}ax_iy_j+\sum_{j=1}^{n}\sum_{i=1}^{n} A_{i,j}x'_iy_j\\ &= a\beta_A(x,y)+\beta_A(x',y)\end{aligned}$$

and

$$\begin{aligned}\beta_A(x,ay+y') &= \sum_{j=1}^{n}\sum_{i=1}^{n} A_{i,j}x_i(ay_j+y'_j)\\ &= \sum_{j=1}^{n}\sum_{i=1}^{n} A_{i,j}ax_iy_j+\sum_{j=1}^{n}\sum_{i=1}^{n} A_{i,j}x_iy'_j\\ &= a\beta_A(x,y)+\beta_A(x,y').\end{aligned}$$

□

Let $V^{(2)}$ denote the set of bilinear forms on V.

Lemma. *$V^{(2)}$ is a vector space, with the usual operations of addition and scalar multiplication of functions:*

$$(\beta_1+\beta_2)(u,v)=\beta_1(u,v)+\beta_2(u,v)$$
$$(k\beta)(u,v)=k\beta(u,v)$$

For T an operator on an n-dimensional vector space V and a basis $\{e_1,\dots,e_n\}$ of V, we used an $n\times n$ matrix to provide information about T. We now do the same for bilinear forms.

Definition 10.2 (Matrix of bilinear form)**.** Suppose $\beta\in V^{(2)}$, and $\{e_1,\dots,e_n\}$ is a basis of V. The matrix of β with respect to this basis is the $n\times n$ matrix $\mathscr{M}(\beta)$ whose entries are defined by

$$\mathscr{M}(\beta)_{i,j}=\beta(e_i,e_j).$$

Hence $\mathscr{M}(\beta;\{e_1,\dots,e_n\})$ is the matrix

$$\begin{pmatrix}\beta(e_1,e_1) & \cdots & \beta(e_1,e_n)\\ \vdots & & \vdots\\ \beta(e_n,e_1) & \cdots & \beta(e_n,e_n)\end{pmatrix}.$$

Let $\{e_1,\dots,e_n\}$ be a basis of V. Let $u,v\in V$. There exist $x_i,y_i\in\mathbf{F}$ such that

$$u=\sum_{i=1}^n x_ie_i,\quad v=\sum_{i=1}^n y_ie_i.$$

By linearity in the second slot,

$$\begin{aligned}\beta(u,v)&=\beta\left(u,\sum_{i=1}^n y_ie_i\right)\\ &=\sum_{i=1}^n\beta(u,e_i)y_i\\ &=\begin{pmatrix}\beta(u,e_1) & \cdots & \beta(u,e_n)\end{pmatrix}\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix}.\end{aligned}$$

For $j = 1, \ldots, n$, by linearity in the first slot,

$$\begin{aligned}\beta(u, e_j) &= \beta\left(\sum_{i=1}^{n} x_i e_i, e_j\right) \\ &= \sum_{i=1}^{n} x_i \beta(e_i, e_j) \\ &= \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix} \begin{pmatrix} \beta(e_1, e_j) \\ \vdots \\ \beta(e_n, e_j) \end{pmatrix}.\end{aligned}$$

Note that

$$\begin{aligned}\begin{pmatrix} \beta(u, e_1) & \cdots & \beta(u, e_n) \end{pmatrix} &= \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix} \begin{pmatrix} \beta(e_1, e_1) & \cdots & \beta(e_1, e_n) \\ \vdots & & \vdots \\ \beta(e_n, e_1) & \cdots & \beta(e_n, e_n) \end{pmatrix} \\ &= \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix} \mathscr{M}(\beta; \{e_1, \ldots, e_n\}).\end{aligned}$$

Therefore

$$\beta(u, v) = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix} \mathscr{M}(\beta) \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

In particular, consider bilinear forms over $\mathbf{F}^n$. Let β be a bilinear form on $\mathbf{F}^n$. Pick the standard basis $\{e_1, \ldots, e_n\}$, where

$$e_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}.$$

Let $x, y \in \mathbf{F}^n$. Then

$$\beta(x, y) = x^\top A y \tag{10.1}$$

where $A = \mathscr{M}(\beta; \{e_1, \ldots, e_n\})$.

Lemma 10.3. $\dim V^{(2)} = (\dim V)^2$.

Proof. Let $\{e_1, \ldots, e_n\}$ be a basis of V.

Consider the map which sends a bilinear form to its matrix.

Claim. The map $\begin{aligned} \mathscr{M} : V^{(2)} &\to \mathscr{M}_{n\times n}(\mathbf{F}) \\ \beta &\mapsto \mathscr{M}(\beta) \end{aligned}$ is an isomorphism.

- $\mathscr{M}$ is a linear map: for all $\beta_1, \beta_2 \in V^{(2)}$,

$$\begin{aligned}\mathscr{M}(\beta_1+\beta_2) &= \big((\beta_1+\beta_2)(e_i,e_j)\big)_{n\times n}\\ &= \big(\beta_1(e_i,e_j)\big)_{n\times n} + \big(\beta_2(e_i,e_j)\big)_{n\times n}\\ &= \mathscr{M}(\beta_1)+\mathscr{M}(\beta_2).\end{aligned}$$

  Similarly,

$$\mathscr{M}(k\beta) = k\mathscr{M}(\beta_i).$$

- $\mathscr{M}$ is injective: let $\beta, \beta' \in V^{(2)}$,

$$\begin{aligned}&\mathscr{M}(\beta) = \mathscr{M}(\beta')\\ \Longrightarrow\ &\beta(e_i,e_j) = \beta'(e_i,e_j) \quad (\forall i,j)\\ \Longrightarrow\ &\beta(u,v) = \beta'(u,v) \quad (\forall u,v \in V)\end{aligned}$$

  since u, v are linear combinations of $e_1, \ldots, e_n$.

  For all $u, v \in V$, let $u = \sum_{i=1}^n x_i e_i$, $v = \sum_{i=1}^n y_i e_i$, $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$. Then

$$\beta(u,v) = x^\top \mathscr{M}(\beta) y = x^\top \mathscr{M}(\beta') y = \beta'(u,v).$$

  Thus this implies $\beta = \beta'$.

- $\mathscr{M}$ is surjective: for all $A \in \mathscr{M}_{n\times n}(\mathbf{F})$, we want to find $\beta \in V^{(2)}$ such that $\mathscr{M}(\beta) = A$, i.e., $\beta(e_i, e_j) = A_{i,j}$.

  For all $u, v \in V$, if $= \sum_{i=1}^n x_i e_i$, $v = \sum_{i=1}^n y_i e_i$, then define $\beta : V \times V \to \mathbf{F}$ by

$$\beta(u,v) = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix} A_{i,j} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

Fix $v \in V$, then for all $u_1, u_2 \in V$, let $u_1 = \sum_{i=1}^n x_i^{(1)} e_i$, $u_2 = \sum_{i=1}^n x_i^{(2)} e_i$. Then

$$\begin{aligned}
\beta(u_1+u_2, v) &= \beta\left(\sum_{i=1}^n \left(x_i^{(1)} + x_i^{(2)}\right) e_i, v\right) \\
&= \begin{pmatrix} x_1^{(1)} + x_1^{(2)} & \cdots & x_n^{(1)} + x_n^{(2)} \end{pmatrix} A_{i,j} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\
&= \begin{pmatrix} x_1^{(1)} & \cdots & x_n^{(1)} \end{pmatrix} A_{i,j} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} + \begin{pmatrix} x_1^{(2)} & \cdots & x_n^{(2)} \end{pmatrix} A_{i,j} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\
&= \beta(u_1, v) + \beta(u_2, v).
\end{aligned}$$

Hence $\beta \in V^{(2)}$ such that $\mathscr{M}(\beta) = A$.

Therefore $\mathscr{M} : V^{(2)} \to \mathscr{M}_{n\times n}(\mathbf{F})$ is an isomorphism. In particular,

$$\dim V^{(2)} = \dim \mathscr{M}_{n\times n} = n^2 = (\dim V)^2.$$

□

Remark. The map $\mathscr{M}$ is not canonical, since it depends on the basis chosen.

The next result allows us to compute the matrix of the composition of a bilinear form and an operator.

Lemma 10.4. *Suppose β is a bilinear form on V and $T \in \mathscr{L}(V)$. Define bilinear forms α and ρ on V by*

$$\alpha(u,v) = \beta(u, Tv), \quad \rho(u,v) = \beta(Tu, v).$$

Let $\{e_1, \ldots, e_n\}$ be a basis of V. Then

$$\mathscr{M}(\alpha) = \mathscr{M}(\beta)\mathscr{M}(T), \quad \mathscr{M}(\rho) = \mathscr{M}(T)^\top \mathscr{M}(\beta).$$

Proof. If $i, j \in \{1, \ldots, n\}$, then

$$\begin{aligned}
\mathscr{M}(\alpha)_{i,j} &= \alpha(e_i, e_j) \\
&= \beta(e_i, Te_j) \\
&= \beta\left(e_i, \sum_{k=1}^n \mathscr{M}(T)_{k,j} e_k\right) \\
&= \sum_{k=1}^n \beta(e_i, e_k) \mathscr{M}(T)_{k,j} \\
&= \left(\mathscr{M}(\beta)\mathscr{M}(T)\right)_{i,j}.
\end{aligned}$$

Hence $\mathscr{M}(\alpha) = \mathscr{M}(\beta)\mathscr{M}(T)$. The proof that $\mathscr{M}(\rho) = \mathscr{M}(T)^\top \mathscr{M}(\beta)$ is similar. □

The result below shows how the matrix of a bilinear form changes if we change the basis.

Proposition 10.5 (Change-of-basis formula). *Suppose $\beta \in V^{(2)}$. Suppose $\{e_1,\dots,e_n\}$ and $\{f_1,\dots,f_n\}$ are bases of V. Let*

$$A = \mathscr{M}(\beta;\{e_1,\dots,e_n\}), \quad B = \mathscr{M}(\beta;\{f_1,\dots,f_n\})$$

and $C = \mathscr{M}(I;\{e_1,\dots,e_n\},\{f_1,\dots,f_n\})$. Then

$$A = C^\top BC.$$

Proof. By the linear map lemma, there exists an operator $T \in \mathscr{L}(V)$ such that

$$Tf_i = e_i \quad (i = 1,\dots,n).$$

THe definition of the matrix of an operator with respect to a basis implies that

$$\mathscr{M}(T;\{f_1,\dots,f_n\}) = C.$$

Define bilinear forms α,ρ on V by

$$\begin{aligned}\alpha(u,v) &= \beta(u,Tv)\\ \rho(u,v) &= \alpha(Tu,v) = \beta(Tu,Tv).\end{aligned}$$

Then $\beta(e_i,e_j) = \beta(Tf_i,Tf_j) = \rho(f_i,f_j)$ for all $i,j \in \{1,\dots,n\}$. Thus

$$\begin{aligned}A &= \mathscr{M}(\rho;\{f_1,\dots,f_n\})\\ &= C^\top \mathscr{M}(\alpha;\{f_1,\dots,f_n\})\\ &= C^\top BC.\end{aligned}$$

□

Symmetric Bilinear Forms

Definition 10.6 (Symmetric bilinear form). We say $\rho \in V^{(2)}$ is ***symmetric*** if

$$\rho(u,w) = \rho(w,u) \quad (u,w \in V).$$

Example.

- The inner product defined on a real inner product space is a symmetric bilinear form:
$$\rho(u,w) = \langle u,w\rangle .$$
- Suppose V is a real inner product space and $T \in \mathscr{L}(V)$. Define $\rho \in V^{(2)}$ by
$$\rho(u,w) = \langle u,Tw\rangle .$$
Then ρ is symmetric bilinear form on V if and only if T is self-adjoint.
- Suppose $\rho : \mathscr{L}(V) \times \mathscr{L}(V) \to \mathbf{F}$ is defined by
$$\rho(S,T) = \operatorname{tr}(ST).$$
Then ρ is a symmetric bilinear form on $\mathscr{L}(V)$, because trace is a linear functional on $\mathscr{L}(V)$, and $\operatorname{tr}(ST) = \operatorname{tr}(TS)$ for all $S,T \in \mathscr{L}(V)$.

The set of symmetric bilinear forms on V is denoted by $V^{(2)}_{\text{sym}}$.

Lemma. *$V^{(2)}_{sym}$ is a subspace of $V^{(2)}$.*

The next result provides a characterisation of symmetric bilinear forms.

Lemma 10.7. *Suppose $\rho \in V^{(2)}$. Then the following are equivalent:*

(i) $\rho \in V^{(2)}_{sym}$.

(ii) $\mathscr{M}(\rho;\{e_1,\dots,e_n\})$ *is a symmetric matrix for every basis* $\{e_1,\dots,e_n\}$ *of* V.

(iii) $\mathscr{M}(\rho;\{e_1,\dots,e_n\})$ *is a symmetric matrix for some basis* $\{e_1,\dots,e_n\}$ *of* V.

(iv) $\mathscr{M}(\rho;\{e_1,\dots,e_n\})$ *is a diagonal matrix for some basis* $\{e_1,\dots,e_n\}$ *of* V.

(ii) and (iii) imply that a bilinear form on V has a symmetric matrix with respect to either *all* bases of V, or *no* bases of V.

Proof.

$\boxed{\text{(i)} \implies \text{(ii)}}$ Suppose ρ is a symmetric bilinear form. Let $\{e_1,\dots,e_n\}$ is a basis of V. If $i,j \in \{1,\dots,n\}$, since ρ is symmetric,

$$\mathscr{M}(\beta)_{i,j} = \rho(e_i,e_j) = \rho(e_j,e_i) = \mathscr{M}(\beta)_{j,i}.$$

Hence $\mathscr{M}(\rho;\{e_1,\dots,e_n\})$ is a symmetric matrix.

$\boxed{\text{(ii)} \Longrightarrow \text{(iii)}}$ This is clear.

$\boxed{\text{(iii)} \Longrightarrow \text{(i)}}$ Suppose $\{e_1, \ldots, e_n\}$ is a basis of V such that $\mathscr{M}(\rho; \{e_1, \ldots, e_n\})$ is a symmetric matrix.

Let $u, w \in V$. Then $u = \sum_{i=1}^{n} a_i e_i$, $w = \sum_{j=1}^{n} b_j e_j$ for some $a_i, b_i \in \mathbf{F}$. Now

$$\begin{aligned}
\rho(u,w) &= \rho\left(\sum_{i=1}^{n} a_i e_i, \sum_{j=1}^{n} b_j e_j\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j \rho(e_i, e_j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j \rho(e_j, e_i) \\
&= \rho\left(\sum_{j=1}^{n} b_j e_j, \sum_{i=1}^{n} a_i e_i\right) \\
&= \rho(w,u).
\end{aligned}$$

$\boxed{\text{(iv)} \Longrightarrow \text{(iii)}}$ This follows since every diagonal matrix is symmetric.

$\boxed{\text{(i)} \Longrightarrow \text{(iv)}}$ Induct on $n = \dim V$.

If $n = 1$, then every 1×1 matrix i diagonal.

Suppose $n > 1$ and the implication (i) $\Longrightarrow$ (iv) holds for one less dimension.

Suppose ρ is a symmetric bilinear form. If $\rho = 0$, then the matrix of ρ with respect to every basis of V is the zero matrix, which is a diagonal matrix. Hence assume $\rho \neq 0$, which means there exist $u, w \in V$ such that $\rho(u,w) \neq 0$. Now

$$2\rho(u,w) = \rho(u+w, u+w) - \rho(u,u) - \rho(w,w).$$

Since the LHS is non-zero, the three terms on the RHS cannot all equal 0. Hence there exists $v \in V$ such that $\rho(v,v) \neq 0$.

Let $U = \{u \in V \mid \rho(u,v) = 0\}$. Thus U is the null space of the linear functional $u \mapsto \rho(u,v)$ on V. This linear functional is not the zero linear functional because $v \notin U$. Thus $\dim U = n-1$. By induction hypothesis, there is a basis $\{e_1, \ldots, e_{n-1}\}$ of U such that the symmetric bilinear form $\rho|_{U \times U}$ has a diagonal matrix with respect to this basis.

Since $v \notin U$, the set $\{e_1, \ldots, e_{n-1}, v\}$ is a basis of V. Let $i \in \{1, \ldots, n-1\}$. Then $\rho(e_i, v) = 0$ by the construction of U. Since ρ is symmetric, we also have $\rho(v, e_i) = 0$. Thus the matrix of ρ with respect to $\{e_1, \ldots, e_{n-1}, v\}$ is a diagonal matrix. □

The previous result states that every symmetric bilinear form has a diagonal matrix with respect to some basis. If our vector space happens to be a real inner product space, then the next result shows that every symmetric bilinear form has a diagonal matrix with respect to some

orthonormal basis.

Proposition 10.8. *Suppose V is a real inner product space, and $\rho \in V^{(2)}_{sym}$. Then ρ has a diagonal matrix with respect to some orthonormal basis of V.*

Proof. Let $\{f_1, \dots, f_n\}$ be an orthonormal basis of V. Let $B = \mathscr{M}(\rho; \{f_1, \dots, f_n\})$. Since $\rho \in V^{(2)}_{\text{sym}}$, by 10.7, B is a symmetric matrix.

Let $T \in \mathscr{L}(V)$ be the operator such that $\mathscr{M}(T; \{f_1, \dots, f_n\}) = B$. Then

$$\mathscr{M}(T^*; \{f_1, \dots, f_n\}) = B^\top = B = \mathscr{M}(T; \{f_1, \dots, f_n\}),$$

so T is self-adjoint. By the real spectral theorem (8.19), T has a diagonal matrix with respect to some orthonormal basis $\{e_1, \dots, e_n\}$ of V.

Let $C = \mathscr{M}(I; \{e_1, \dots, e_n\}, \{f_1, \dots, f_n\})$. By the change-of-basis formula (4.46),

$$\mathscr{M}(T; \{e_1, \dots, e_n\}) = C^{-1}BC.$$

Hence $C^{-1}BC$ is a diagonal matrix. Now

$$\mathscr{M}(\rho; \{e_1, \dots, e_n\}) = C^\top BC = C^{-1}BC,$$

where the first equality holds by the change-of-basis formula for bilinear forms (10.5), and the second equality holds because C is a unitary matrix with real entries (which implies that $C^{-1} = C^\top$; see 8.33).

Therefore ρ has a diagonal matrix with respect to $\{e_1, \dots, e_n\}$. □

Alternating Bilinear Forms

Definition 10.9 (Alternating bilinear form)**.** We say $\alpha \in V^{(2)}$ is ***alternating*** if

$$\alpha(v, v) = 0 \quad (v \in V).$$

The set of alternating bilinear forms on V is denoted by $V^{(2)}_{\text{alt}}$.

Lemma. *$V^{(2)}_{alt}$ is a subspace of V.*

The next result shows that a bilinear form is alternating if and only if switching the order of the two inputs multiplies the output by -1.

Lemma 10.10. *$\alpha \in V_{alt}^{(2)}$ if and only if*

$$\alpha(u,w) = -\alpha(w,u) \quad (u,w \in V).$$

Proof.

$\boxed{\Longrightarrow}$ Suppose $\alpha \in V_{\text{alt}}^{(2)}$. Let $u,w \in V$, then

$$\begin{aligned} 0 &= \alpha(u+w,u+w) \\ &= \alpha(u,u) + \alpha(u,w) + \alpha(w,u) + \alpha(w,w) \\ &= \alpha(u,w) + \alpha(w,u). \end{aligned}$$

Thus $\alpha(u,w) = -\alpha(w,u)$.

$\boxed{\Longleftarrow}$ Suppose $\alpha(u,w) = -\alpha(w,u)$ for all $u,w \in V$. Then $\alpha(v,v) = -\alpha(v,v)$ for all $v \in V$, which implies $\alpha(v,v) = 0$ for all $v \in V$. Thus α is alternating. □

Now we show that the vector space of bilinear forms on V is the direct sum of the symmetric bilinear forms on V and the alternating bilinear forms on V.

Proposition 10.11.

$$V^{(2)} = V_{sym}^{(2)} \oplus V_{alt}^{(2)}.$$

Proof. We first show $V^{(2)} = V_{\text{sym}}^{(2)} + V_{\text{alt}}^{(2)}$. Let $\beta \in V^{(2)}$. Define $\rho, \alpha \in V^{(2)}$ by

$$\rho(u,w) = \frac{\beta(u,w) + \beta(w,u)}{2}, \quad \alpha(u,w) = \frac{\beta(u,w) - \beta(w,u)}{2}.$$

Then $\rho \in V_{\text{sym}}^{(2)}$ and $\alpha \in V_{\text{alt}}^{(2)}$, and $\beta = \rho + \alpha$. Hence $V^{(2)} = V_{\text{sym}}^{(2)} + V_{\text{alt}}^{(2)}$.

Next we show the direct sum. Let $\beta \in V_{\text{sym}}^{(2)} \cap V_{\text{alt}}^{(2)}$. If $u,w \in V$,

$$\beta(u,w) = -\beta(w,u) = -\beta(u,w)$$

so $\beta(u,w) = 0$. Hence $\beta = 0$. □

Quadratic Forms

Definition 10.12 (Quadratic form). We say $q\colon V \to \mathbf{F}$ is a ***quadratic form*** on V if there exists $\beta \in V^{(2)}$ such that

$$q(v) = \beta(v,v) \quad (v \in V).$$

The next result characterises quadratic forms on $\mathbf{F}^n$.

Lemma 10.13. *$q\colon \mathbf{F}^n \to \mathbf{F}$ is a quadratic form on $\mathbf{F}^n$ if and only if there exist $A_{i,j} \in \mathbf{F}$ such that*

$$q(x_1,\dots,x_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{i,j}x_ix_j$$

for all $(x_1,\dots,x_n) \in \mathbf{F}^n$.

Proof. Let $\beta \in V^{(2)}$ be such that $q(v) = \beta(v,v)$. Pick the standard basis $\{e_1,\dots,e_n\}$ of $\mathbf{F}^n$. Then

$$\begin{aligned} q(x_1,\dots,x_n) &= q(x_1e_1+\cdots+x_ne_n) \\ &= \beta(x_1e_1+\cdots+x_ne_n, y_1e_1+\cdots+y_ne_n) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} x_ix_j\beta(e_i,e_j). \end{aligned}$$

Pick $A = \mathscr{M}(\beta;\{e_1,\dots,e_n\})$. □

The next result provides characterisations of quadratic forms.

Although quadratic forms are defined in terms of an arbitrary bilinear form, the equivalence of (i) and (ii) shows that a *symmetric* bilinear form can always be used.

Lemma 10.14. *Suppose $q\colon V \to \mathbf{F}$. Then the following are equivalent:*

(i) q is a quadratic form.

(ii) There exists a unique $\rho \in V^{(2)}_{sym}$ such that $q(v,v) = \rho(v,v)$.

(iii) $q(\lambda v) = \lambda^2 q(v)$ for all $v \in V$, $\lambda \in \mathbf{F}$, and the function

$$(u,w) \mapsto q(u+w) - q(u) - q(w)$$

is a symmetric bilinear form on V.

(iv) $q(2v) = 4q(v)$ for all $v \in V$, and the function

$$(u,w) \mapsto q(u+w) - q(u) - q(w)$$

is a symmetric bilinear form on V.

Proof.

(i) $\Longrightarrow$ (ii) Suppose q is a quadratic form. Then $q(v) = \beta(v,v)$ for some $\beta \in V^{(2)}$.

Claim. $\rho(v,w) = \dfrac{\beta(v,w)+\beta(w,v)}{2}$, for all $v,w \in V$.

It is easy to check that $\rho \in V^{(2)}_{\text{sym}}$, and $q(v) = \beta(v,v) = \rho(v,v)$ for all $v \in V$.

To prove uniqueness, suppose $\rho' \in V_{\text{sym}}^{(2)}$ is such that $q(v) = \rho'(v,v)$. Then $(\rho' - \rho)(v,v) = 0$ for all $v \in V$, or $\rho' - \rho = 0$. Thus $\rho' = \rho$.

(ii) $\Longrightarrow$ (iii) Suppose there exists $\rho \in V_{\text{sym}}^{(2)}$ such that $q(v,v) = \rho(v,v)$. Let $\lambda \in \mathbf{F}$, $v \in V$, then

$$q(\lambda v) = \rho(\lambda v, \lambda v) = \lambda \rho(v, \lambda v) = \lambda^2 \rho(v,v) = \lambda^2 q(v).$$

Let $u, w \in V$, then

$$q(u+w) - q(u) - q(w) = \rho(u+w, u+w) - \rho(u,u) - \rho(w,w) = 2\rho(u,w).$$

Thus the function $(u,w) \mapsto q(u+w) - q(u) - q(w)$ equals 2ρ, which is a symmetric bilinear form on V.

(iii) $\Longrightarrow$ (iv) This is obvious.

(iv) $\Longrightarrow$ (i) Let $\rho \in V_{\text{sym}}^{(2)}$ be defined by

$$\rho(u,w) = \frac{q(u+w) - q(u) - q(w)}{2}.$$

Let $v \in V$, then

$$\rho(v,v) = \frac{q(2v) - q(v) - q(v)}{2} = \frac{4q(v) - 2q(v)}{2} = q(v).$$

Hence q is a quadratic form. □

Proposition 10.15. *Suppose q is a quadratic form on V. Then there exists a basis $\{e_1, \ldots, e_n\}$ of V, and $\lambda_1, \ldots, \lambda_n \in \mathbf{F}$ such that*

$$q(x_1 e_1 + \cdots + x_n e_n) = \lambda_1 {x_1}^2 + \cdots + \lambda_n {x_n}^2$$

for all $x_1, \ldots, x_n \in \mathbf{F}$.

If V is a real inner product space, by 10.8, the basis $\{e_1, \ldots, e_n\}$ can be chosen to be an orthonormal basis of V.

Proof. Since q is a quadratic form on V, by 10.14, there exists $\rho \in V_{\text{sym}}^{(2)}$ such that $q(v) = \rho(v,v)$. By 10.7, there exists a basis $\{e_1, \ldots, e_n\}$ of V such that $\mathscr{M}(\rho; \{e_1, \ldots, e_n\})$ is a diagonal matrix. Let $\lambda_1, \ldots, \lambda_n$ denote the entries on its diagonal. Thus for all $i, j \in \{1, \ldots, n\}$,

$$\rho(e_i, e_j) = \begin{cases} \lambda_i & (i = j) \\ 0 & (i \neq j) \end{cases}$$

If $x_1, \ldots, x_n \in \mathbf{F}$, then

$$\begin{aligned} q(x_1e_1 + \cdots + x_ne_n) &= \rho(x_1e_1 + \cdots + x_ne_n, x_1e_1 + \cdots + x_ne_n) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j \rho(e_i, e_j) \\ &= \lambda_1 {x_1}^2 + \cdots + \lambda_n {x_n}^2. \end{aligned}$$

□

10.2 Alternating Multilinear Forms

Multilinear Forms

m-linear forms are a generalisation of bilinear forms.

Definition 10.16 (Multilinear form). For m a positive integer, an ***m-linear form*** on V is a function $\beta : V^m \to \mathbf{F}$ that is linear in each slot when the other slots are held fixed: for each $i \in \{1,\dots,m\}$ and all $u_1,\dots,u_m \in V$, the function

$$v \mapsto \beta(u_1,\dots,u_{i-1},v,u_{i+1},\dots,u_m)$$

is a linear map from V to $\mathbf{F}$.
We say $\beta : V^m \to \mathbf{F}$ is a ***multilinear form*** on V if it is an m-linear form on V for some positive integer m.

The set of m-linear forms on V is denoted by $V^{(m)}$.

Definition 10.17 (Alternating multilinear form). We say $\alpha \in V^{(m)}$ is ***alternating*** if

$$\alpha(v_1,\dots,v_m) = 0$$

whenever $v_1,\dots,v_m \in V$ with $v_i = v_j$ for some distinct $i,j \in \{1,\dots,m\}$.

The set of alternating m-linear forms on V is denoted by $V^{(m)}_{\text{alt}}$.

Lemma. *$V^{(m)}_{alt}$ is a subspace of $V^{(m)}$.*

The next result tells us that if a linearly dependent list is input to an alternating multilinear form, then the output equals 0.

Lemma 10.18. *Suppose $\alpha \in V^{(m)}_{alt}$. If $v_1,\dots,v_m$ are linearly dependent in V, then*

$$\alpha(v_1,\dots,v_m) = 0.$$

The next result states that if $m > \dim V$, then there are no alternating m-linear forms on V other than the function on V^m that is identically 0.

Lemma 10.19. *Suppose $m > \dim V$. Then 0 is the only alternating m-linear form on V.*

Alternating Multilinear Forms and Permutations

Lemma 10.20. *Suppose $\alpha \in V_{alt}^{(m)}$, and $v_1, \ldots, v_m \in V$. Then swapping the vectors in any two slots of $\alpha(v_1, \ldots, v_m)$ changes the value of α by a factor of -1.*

To deal with arbitrary multiple swaps, we need a bit of information about permutations.

Definition 10.21 (Permutation)**.**

Define an *inversion* to be a pair of elements that are out of their natural order. That is, if $1 \leq k < l \leq m$, then (k,l) is an inversion if k appears after l in the list $(i_1, \ldots, i_m)$.

Definition 10.22 (Sign of permutation)**.** The ***sign*** of a permutation $(i_1, \ldots, i_m)$ is defined by

$$\operatorname{sgn}(i_1, \ldots, i_m) := (-1)^N$$

where N is the number of inversions in $(i_1, \ldots, i_m)$.

Hence the sign of a permutation equals 1 if the natural order has been changed an even number of times, and equals -1 if the natural order has been changed an odd number of times.

Lemma 10.23. *Swapping two entries in a permutation multiplies the sign of the permutation by -1.*

Lemma 10.24.

Our use of permutations now leads in a natural way to the following beautiful formula for alternating n-linear forms on an n-dimensional vector space.

Lemma 10.25. *Let $n = \dim V$. Suppose $\{e_1, \ldots, e_n\}$ is a basis of V, and $v_1, \ldots, v_n \in V$. For each $b_{1,k}, \ldots, b_{n,k} \in \mathbf{F}$ be such that*

$$v_k = \sum_{i=1}^{n} b_{i,k} e_i.$$

Then for every $\alpha \in V_{alt}^{(n)}$,

The following result will be the key to our definition of the determinant in the next section.

Proposition 10.26. $\dim V_{alt}^{(\dim V)} = 1$.

Lemma 10.27. *Let $n = \dim V$. Suppose $\alpha \in V_{alt}^{(n)}$ is non-zero, and $e_1, \ldots, e_n \in V$. Then*

$$\alpha(e_1, \ldots, e_n) \neq 0$$

if and only if $e_1, \ldots, e_n$ are linearly independent.

10.3 Determinants

Defining the Determinant

Properties of Determinants

10.4 Tensor Products

Tensor Product of Two Vector Spaces

Tensor Product of Inner Product Spaces

Tensor Product of Multiple Vector Spaces

Exercises

Exercise 10.1 ([Axl24] 9A Q1)**.** Prove that if β is a bilinear form on $\mathbf{F}$, then there exists $c \in \mathbf{F}$ such that

$$\beta(x,y) = cxy$$

for all $x, y \in \mathbf{F}$.

Solution. We note that since the input is taken from $\mathbf{F}$, the basis is naturally 1. Thus we have

$$\beta(x,y) = x\beta(1,y) = xy\beta(1,1) = cxy$$

where we take $c = \beta(1,1)$. □

Exercise 10.2 ([Axl24] 9A Q2)**.** Let $n = \dim V$. Suppose β is a bilinear form on V. Prove that there exist $\phi_1, \ldots, \phi_n, \tau_1, \ldots, \tau_n \in V'$ such that

$$\beta(u,v) = \phi_1(u) \cdot \tau_1(v) + \cdots + \phi_n(u) \cdot \tau_n(v)$$

for all $u, v \in V$.

Solution. Let $\{e_1, \ldots, e_n\}$ be a basis of V, with dual basis $\{e_1^*, \ldots, e_n^*\}$. We have

$$\begin{aligned}
\beta(u,v) &= \beta(x_1e_1 + \cdots + x_ne_n, y_1e_1 + \cdots + y_ne_n) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} x_iy_j\beta(e_i,e_j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \beta(e_i,e_j)e_i^*(u)e_j^*(v) \quad [x_i = e_i^*(u), y_i = e_i^*(v)] \\
&= \sum_{j=1}^{n}\left(\sum_{i=1}^{n}\beta(e_i,e_j)e_i^*(u)\right)e_j^*(v) \\
&= \sum_{j=1}^{n}\left(\sum_{i=1}^{n}\beta(e_i,e_j)e_i^*\right)(u)e_j^*(v).
\end{aligned}$$

Let

$$\begin{aligned}
\phi_j &= \sum_{i=1}^{n}\beta(e_i,e_j)e_i^* \in V' \\
\tau_j &= e_j^* \in V'
\end{aligned}$$

Then $\beta(u,v) = \sum_{j=1}^{n}\phi_j(u)\tau_j(v)$. □

Exercise 10.3 ([Axl24] 9A Q4)**.** Suppose V is a real inner product space and β is a bilinear form on V. Show that there exists a unique operator $T \in \mathscr{L}(V)$ such that

$$\beta(u,v) = \langle u, Tv \rangle$$

for all $u, v \in V$.

Solution. Pick an orthonormal basis $\{e_1, \ldots, e_n\}$ of V. □

Exercise 10.4 ([Axl24] 9A Q5)**.** Suppose β is a bilinear form on a real inner product space V and T is the unique operator on V such that $\beta(u,v) = \langle u, Tv \rangle$ for all $u, v \in V$ (see Exercise 4). Show that β is an inner product on V if and only if T is an invertible positive operator on V.

Solution.

$\boxed{\Longleftarrow}$ We check that β is an inner product on V:

(i) Since β is bilinear, this implies β is sesquilinear.

(ii) Since T is self-adjoint, $\beta(v,u) = \langle v, Tu \rangle = \langle T^*v, u \rangle = \langle Tv, u \rangle = \langle u, Tv \rangle = \beta(u,v)$. Hence β is symmetric.

(iii) Since T is a positive operator, it has a square root $\sqrt{T}$. Thus $\beta(v,v) = \langle v, Tv \rangle = \left\langle \sqrt{T}v, \sqrt{T}v \right\rangle = \left\| \sqrt{T}v \right\|^2 \geq 0$. Also $\beta(v,v) = 0 \iff \sqrt{T}v = \mathbf{0} \iff Tv = \mathbf{0} \iff v = \mathbf{0}$ since T is invertible. Hence positive definiteness holds.

$\boxed{\Longrightarrow}$ T is self-adjoint: for all $u, v \in V$,

$$\langle Tu, v \rangle = \langle v, Tu \rangle = \beta(v,u) = \beta(u,v) = \langle u, Tv \rangle = \langle T^*u, v \rangle .$$

Thus $Tu = T^*u$ for all $u \in V$, so $T^* = T$.

T is positive:

$$\langle Tv, v \rangle = \langle v, Tv \rangle = \beta(v,v) \geq 0$$

by positive definiteness of inner product.

T is invertible: Suppose $Tv = \mathbf{0}$. Then for all $u \in V$,

$$\beta(u,v) = \langle u, Tv \rangle = 0.$$

Pick $u = v$. By positive definiteness of inner product, $\beta(v,v) = \langle v, Tv \rangle = 0$. Hence $v = \mathbf{0}$. □

Exercise 10.5 ([Axl24] 9A Q8)**.** Find formulae for $\dim V^{(2)}_{\text{sym}}$ and $\dim V^{(2)}_{\text{alt}}$ in terms of $\dim V$.

Solution. Let $n = \dim V$. For $\beta \in V^{(2)}_{\text{sym}}$, consider $\mathscr{M}(\beta)$. Its diagonal entries can be chosen arbitrarily. For off-diagonal entries, only half of them can be chosen arbitrarily, therefore the dimension is $\frac{n(n-1)}{2} + n = \frac{n(n+1)}{2}$.

For $\beta \in V^{(2)}_{\text{alt}}$, consider $\mathscr{M}(\beta)$. The diagonal entires are all 0 and only half of the off-diagonal entries can be chosen arbitrarily. Therefore, the dimension is $\frac{n(n-1)}{2}$. □

III
Abstract Algebra

Algebra is the study of collections of objects (sets, groups, rings, fields, etc). In algebra, we are concerned about the structures of these collections and how these collections interact than about the objects themselves. In fact, with homomorphism and isomorphisms, the original objects become irrelevant.

11 Groups

One of the simplest forms of abstract algebraic systems is a *group*, which is roughly a set of objects and a rule for multiplying them together. Groups arise all over mathematics, particularly where there is symmetry.

11.1 Groups

Definitions and Properties

A ***binary operation*** on a set G is a map $*\colon G \times G \to G$.

Notation. For any $a, b \in G$, if the operation is clear, we write ab for the image of (a,b) under $*$.

Definition 11.1 (Group). A ***group*** $(G,*)$ consists of a set G and a binary operation $*$ on G satisfying the following properties:

(i) $a(bc) = (ab)c$ for all $a,b,c \in G$; (associativity)

(ii) there exists $e \in G$ such that $ae = ea = a$ for all $a \in G$; (identity)

(iii) for all $a \in G$, there exists $c \in G$ such that $ac = ca = e$. (invertibility)

Notation. If the operation is clear, we simply denote a group $(G,*)$ by G.

For the rest of this text, G denotes a group.

Remark. When verifying that G is a group we have to check (i), (ii), (iii) above, and also that $*$ is a binary operation closed in G: $ab \in G$ for all $a,b \in G$.

Notation. Since $*$ is associative, we omit unnecessary parentheses and write $(ab)c = a(bc) = abc$.

We say G is ***abelian*** if the operation is commutative; otherwise, G is *non-abelian*.

Lemma 11.2. *A group has a unique identity.*

Proof. Suppose that e and e' are identities of G. Then

$$e = ee' = e'$$

where the first equality holds since e' is an identity, and the second equality holds since e is an identity. □

Lemma 11.3. *Each element of a group has a unique inverse.*

Proof. Suppose that b and c are both inverses of a. Then $ab = e$, $ca = e$, so

$$c = ce = c(ab) = (ca)b = eb = b.$$

□

We denote *the* inverse of $a \in G$ as a^{-1}.

Lemma 11.4.

(i) $(a^{-1})^{-1} = a$ *for all* $a \in G$.

(ii) $(ab)^{-1} = b^{-1}a^{-1}$ *for all* $a, b \in G$.

(iii) For any $a_1, \ldots, a_n \in G$, $a_1 \cdots a_n$ *is independent of how we arrange the parantheses (generalised associative law).*

Proof.

(i) To show $(a^{-1})^{-1} = a$ is exactly the problem of showing that a is the inverse of a^{-1}, which is by definition of the inverse (with the roles of a and a^{-1} interchanged).

(ii) Let $c = (ab)^{-1}$. Then $(ab)c = e$, or $a(bc) = e$ by associativity, which gives $bc = a^{-1}$. Applying b^{-1} on both sides gives $c = b^{-1}a^{-1}$.

(iii) Induct on n. The result is trivial for $n = 1, 2, 3$. For all $k < n$ assume that any $a_1 \cdots a_k$ is independent of parantheses. Then

$$(a_1 \cdots a_n) = (a_1 \cdots a_k)(a_{k+1} \cdots a_n).$$

By inductive hypothesis, both terms are independent of parentheses since $k, n-k < n$. Hence by induction we are done.

□

Lemma 11.5 (Cancellation law). *Let $a, b \in G$. Then the equations $ax = b$ and $ya = b$ have unique solutions for $x, y \in G$.*

This means that we can cancel on the left and right.

Proof. To solve $ax = b$, apply a^{-1} on both sides to get $x = a^{-1}b$. The uniqueness of x follows because a^{-1} is unique.

A similar case holds for $ya = b$. □

We now introduce notation for repeated application of the operation on an element.

Notation. For any $a \in G$, $n \in \mathbb{N}$, denote $a^n = \underbrace{a \cdot a \cdots a}_{n \text{ times}}$, $a^0 = 1$, and $a^{-n} = (a^{-1})^n$.

The usual rules of exponents hold true:

$$a^{m+n} = a^m a^n$$

$$(a^m)^n = a^{mn}$$

$$(a^n)^{-1} = (a^{-1})^n$$

Definition 11.6 (Order of a group). The ***order*** of G is its cardinality $|G|$. We say G is a *finite group* if $|G| < \infty$.

One way to represent a finite group is by means of a ***Cayley table***. Let $G = \{e, g_2, g_3, \ldots, g_n\}$. The Cayley table of G is a square grid which contains all the possible products of two elements from G: the product $g_i g_j$ appears in the i-th row and j-th column.

Remark. Note that a group is abelian if and only if its Cayley table is symmetric about the main (top-left to bottom-right) diagonal.

Examples of Groups

Example.

- $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ are abelian groups under addition.
- $\mathbb{Q}^\times = \mathbb{Q} \setminus \{0\}$, $\mathbb{R}^\times = \mathbb{R} \setminus \{0\}$, and $\mathbb{C}^\times = \mathbb{C} \setminus \{0\}$ are groups under multiplication.
- The complex numbers of absolute value 1 form a group under multiplication.
- $\{1, -1\}$ is a group under multiplication.
- $\{1, -1, i, -i\}$ is a group under multiplication.

Example (Modular arithmetic)**.** For $n \in \mathbb{N}$, the set of (congruence classes of) integers modulo n, $\mathbb{Z}/n\mathbb{Z}$, is an abelian group under addition.
For $n \in \mathbb{N}$, $(\mathbb{Z}/n\mathbb{Z})^\times$ is an abelian group under multiplication.

Example (Direct product)**.** Let G, H be groups. The cartesian product $G \times H$ is a group under the operation

$$(g_1, h_1) \cdot (g_2, h_2) = (g_1 g_2, h_1 h_2).$$

We call $G \times H$ the *direct product* of G and H.
One may also take a direct product of a finite number of groups. Thus if $G_1, \ldots, G_n$, we let

$$\prod_{i=1}^{n} G_i = G_1 \times \cdots \times G_n$$

be the set of all n-tuples $(x_1, \ldots, x_n)$ with $x_i \in G_i$. We define multiplication componentwise, and see at once that $G_1 \times \cdots \times G_n$ is a group. If e_i is the identity of G_i, then $(e_1, \ldots, e_n)$ is the identity of the product.

Example (Dihedral groups)**.** The ***dihedral group*** of order $2n$, denoted by D_{2n}, is the group of symmetries (rotations and reflections) of a regular n-gon.

- Let r be the rotation clockwise about the origin by $\frac{2\pi}{n}$ radians.
- Let s be the reflection about the line of symmetry through the first labelled vertex and the origin.

We perform the actions from right to left; for instance, sr means do r then s. This is because we think of them as functions on the vertices of the n-gon.
Notice the following properties of D_{2n}:

(i) There are n rotations, since $e, r, r^2, \ldots, r^{n-1}$ are all distinct and $r^n = e$.

(ii) There are 2 rotations, since we either reflect or do not reflect, and $s^2 = e$.

(iii) Since r^i only fixes 1 if $r^i = e$, and s fixes 1, we must have $s \neq r^i$ for any i. This means that the effect of any reflection cannot be obtained from any form of rotation.

(iv) $sr^i \neq sr^j$ for all $i \neq j$ $(0 \leq i, j \leq n-1)$.

This implies that each element of D_{2n} can be written uniquely as $s^i r^j$ for $i \in \{0, 1\}$, $j \in \{0, \ldots, n-1\}$. Hence

$$D_{2n} = \{1, r, \ldots, r^{n-1}, s, sr, \ldots, sr^{n-1}\}$$

and thus $|D_{2n}| = 2n$.

(v) $rs = sr^{-1}$. Thus $rs = sr^{n-1}$, so s and r do not commute (unless $n = 2$). Hence D_{2n} is not abelian.

(vi) More generally, $r^i s = sr^{-i}$ for all $0 \leq i \leq n-1$.

(Proof: This is true for $i = 1$. Assume it holds for $i < n$. Then $r^{i+1}s = r(r^i s) = rsr^{-i}$. Then $rs = sr^{-1}$ so $rsr^{-i} = sr^{-1}r^{-i} = sr^{-i-1}$.)

Note that for each $n \in \mathbb{N}$, the generators of D_{2n} are r and s, and we have shown that they satisfy $r^n = 1$, $s^2 = 1$, and $rs = sr^{-1}$; these are called *relations*. *Any* other equation involving the generators can be derived from these relations.
Any such collection of generators S and relations $R_1, \ldots, R_m$ for a group G is called a *presentation*, written

$$G = \langle S \mid R_1, \ldots, R_m \rangle.$$

For example,

$$D_{2n} = \left\langle r, s \mid r^n = s^2 = 1, rs = sr^{-1} \right\rangle.$$

We show that D_{2n} is a group.

(i) Since function composition is associative, the operation is associative.

(ii) The identity is e, where the n-gon is left unchanged.

(iii) inverses

Example (Matrix groups). For $n \in \mathbb{N}$, let $GL_n(\mathbf{F})$ be the set of all $n \times n$ invertible matrices whose entries are in $\mathbf{F}$:

$$GL_n(\mathbf{F}) = \{A \in M_{n\times n}(\mathbf{F}) \mid \det(A) \neq 0\}.$$

We show that $GL_n(\mathbf{F})$ is a group under matrix multiplication; $GL_n(\mathbf{F})$ is the ***general linear group*** of degree n. Since $\det AB = \det A \det B$, if $\det A \neq 0$ and $\det B \neq 0$, then $\det AB \neq 0$, so $GL_n(\mathbf{F})$ is closed under matrix multiplication.

(i) Matrix multiplication is associative.

(ii) $\det(A) \neq 0$ if and only if A has an inverse matrix, so each $A \in GL_n(\mathbf{F})$ has an inverse $A^{-1} \in GL_n(\mathbf{F})$ such that

$$AA^{-1} = A^{-1}A = I$$

where I is the $n \times n$ identity matrix.

(iii) Inverse

Example (Quaternion group). The ***Quaternion group*** Q_8 is defined by

$$Q_8 = \{1, -1, i, -i, j, -j, k, -k\}$$

with product $\cdot$ computed as follows:

- $1 \cdot a = a \cdot 1 = a$ for all $a \in Q_8$
- $(-1) \cdot (-1) = 1$
- $(-1) \cdot a = a \cdot (-1) = -a$ for all $a \in Q_8$
- $i \cdot i = j \cdot j = k \cdot k = -1$
- $i \cdot j = k$, $j \cdot i = -k$, $j \cdot k = i$, $k \cdot j = -i$, $k \cdot i = j$, $i \cdot k = -j$

Note that Q_8 is a non-abelian group of order 8.

Example (Group of roots of unity). Let $n \in \mathbb{Z}^+$. Consider the set of roots of unity

$$\mu_n = \{z \in \mathbb{C} \mid z^n = 1\} = \{e^{\frac{2k\pi i}{n}} \mid k = 0, \dots, n-1\}.$$

This forms an abelian group under multiplication, of order n.

Subgroups

When given a set with certain properties, it is natural to consider its subsets that inherit the same properties.

Definition 11.7 (Subgroup). We say a non-empty $H \subset G$ is a ***subgroup*** of G, denoted by $H \leq G$, if H is a group under the restricted operation from G.

Example.

- Every group G has two obvious subgroups: the group G itself, and the *trivial subgroup* $\{1\}$. A subgroup is a *proper subgroup* if it is not one of those two.
- $(\mathbb{Q}, +)$ is a subgroup of $(\mathbb{R}, +)$.
- The group of complex numbers of absolute value 1 is a subgroup of $\mathbb{C}^\times$, under multiplication.
- $\{1, -1\}$ is a subgroup of $\{1, -1, i, -i\}$, under multiplication.

- $\{e, r, r^2, \ldots, r^{n-1}\} \leq D_{2n}$ and $\{e, s\} \leq D_{2n}$.
- $S_m \leq S_n$ for $m \leq n$.
- If $H \leq G$ and $K \leq H$, then $K \leq G$. (Thus the relation "is a subgroup of" is transitive.)

According to the definition, to prove that H is a subgroup of G, we need to make sure H satisfies all group axioms. However, this is often tedious. Instead, there are some simplified criteria to decide whether H is a subgroup.

Lemma 11.8. *Let G be a group. Then $H \leq G$ if and only if*

(i) $e \in H$; *(identity)*

(ii) $ab \in H$ *for all* $a, b \in H$; *(closure)*

(iii) $a^{-1} \in H$ *for all* $a \in H$. *(inverses)*

Humans are lazy, and the test above is still too complicated. We thus come up with an even simpler test:

Lemma 11.9 (Subgroup criterion)**.** *Let G be a group. Then $H \leq G$ if and only if*

(i) $H \neq \emptyset$;

(ii) $ab^{-1} \in H$ *for all* $a, b \in H$.

Proof.

$\boxed{\Longrightarrow}$ If $H \leq G$, then we are done, by definition of subgroup.

$\boxed{\Longleftarrow}$ Check group axioms:

(i) Since $H \neq \emptyset$, there exists $a \in H$. Then $e = aa^{-1} \in H$.

(ii) Since $e \in H$ and $a \in H$, then $a^{-1} = ea^{-1} \in H$.

(iii) For any $a, b \in H$, $a, b^{-1} \in H$. By (ii), $a(b^{-1})^{-1} = ab \in H$.

□

The next result and its corollary show that the intersection of subgroups is a subgroup.

Proposition 11.10. *Let G be a group, $H, K \leq G$. Then $H \cap K \leq G$.*

Proof. Apply the subgroup criterion:

(i) Since $e \in H$ and $e \in K$, then $e \in H \cap K$, so $H \cap K \neq \emptyset$.

(ii) Let $a, b \in H \cap K$. Then $a, b \in H$ and $a, b \in K$. Since $H, K \leq G$, by the subgroup criterion, $ab^{-1} \in H$ and $ab^{-1} \in K$, so $ab^{-1} \in H \cap K$.

□

Corollary 11.11. *Let G be a group, $\{H_i \mid i \in I\}$ is a collection of subgroups of G. Then*

$$\bigcap_{i \in I} H_i \leq G.$$

If $S \subset G$, then there exists a smallest subgroup containing S, namely the intersection of all subgroups containing S:

$$\langle S \rangle = \bigcap_{H \leq G, H \supset S} H.$$

We call $\langle S \rangle$ the *subgroup generated* by S.

By 11.11, $\langle S \rangle$ is a subgroup of G.

Lemma 11.12. *An element of G is in $\langle S \rangle$ if and only if it is a finite product of elements of S and their inverses, possibly repeated.*

Proof. Let $S \subset G$. Define

□

Definition 11.13 (Subgroup generated by subset)**.** Let $S \subset G$ be non-empty. Let H be the set of elements of G consisting of all products $x_1 \cdots x_n$ such that $x_i \in S$ or $x_i^{-1} \in S$ for each i, and also containing the unit element.
We call H the ***subgroup generated*** by S. We also say that S is a set of *generators* of H, and denote

$$H = \langle S \rangle .$$

If the elements $\{x_1, \ldots, x_n\}$ form a set of generators for G, we write

$$G = \langle x_1, \ldots, x_n \rangle .$$

Lemma. *The subgroup generated by a subset is a subgroup.*

Proof. □

Cyclic Groups

We consider the subgroup generated by only one element.

Definition 11.14 (Cyclic subgroup)**.** The ***cyclic subgroup*** H generated by $a \in G$ is the set of all powers of a:

$$H = \langle a \rangle = \{a^n \mid n \in \mathbb{Z}\}.$$

We say that a is a *generator* of H.
We say G is ***cyclic*** if there exists $a \in G$ such that $G = \langle a \rangle$.

We write C_n for the cyclic group of order n:

$$C_n = \langle a \mid a^n = 1 \rangle .$$

Lemma. *$H = \langle a \rangle$ is a subgroup of G.*

Proof.

(i) H contains the identity $1 = a^0$.

(ii) Let $a^n, a^m \in H$. Then $a^m a^n = a^{m+n} \in H$.

(iii) $(a^n)^{-1} = a^{-n} \in H$.

□

Example.

- $\mathbb{Z}$ is cyclic with generator 1 or -1. It is *the* infinite cyclic group.
- The multiplicative group $\{1, -1\}$ is cyclic with generator -1.
- $\mathbb{Z}/n\mathbb{Z}$ is cyclic, with all numbers coprime with n as generators.
- The multiplicative group $\{1, -1\}$ is cyclic of order 2.
- The complex numbers $\{1, i, -1, -i\}$ form a cyclic group of order 4. The number i is a generator.

Remark. A cyclic subgroup may have more than one generator. For example, if a is a generator, then a^{-1} is also a generator:

$$\{a^n \mid n \in \mathbb{Z}\} = \{(a^{-1})^n \mid n \in \mathbb{Z}\}.$$

Lemma 11.15. *Cyclic groups are abelian.*

Proof. Let G be a cyclic group. For $a^i, a^j \in G$, we have $a^i a^j = a^{i+j} = a^j a^i$. □

Proposition 11.16. *A subgroup of a cyclic group is cyclic.*

Proof. Let $a \in G$, $H \leq \langle a \rangle$. If $H = \{1\}$ then trivially H is cyclic.

Suppose that H contains some other element $b \neq 1$. Then $b = a^n$ for some integer n. Since H is a subgroup, $b^{-1} = a^{-n} \in H$. Since either n or $-n$ is positive, we can assume H contains positive powers of a and $n > 0$. Let m be the smallest positive integer such that $a^m \in H$ (such an m exist by the well-ordering principle).

Claim. $h = a^m$ is a generator for H.

We need to show that every $h' \in H$ can be written as a power of h. Since $h' \in H$ and $H \leq \langle a \rangle$, $h' = a^k$ for some integer k. By the division algorithm, there exist integers q, r such that $k = qm + r$ with $0 \leq r < m$. Hence

$$a^k = a^{qm+r} = (a^m)^q a^r = h^q a^r$$

so $a^r = a^k h^{-q}$. Since $a^k, h^{-q} \in H$, we must have $a^r \in H$. By the minimality of m, we must have $m = 0$ and so $k = qm$. Hence

$$h' = a^k = a^{qm} = h^q$$

and H is generated by h. □

This result allows us to determine *all* the subgroups of $\mathbb{Z}$.

Corollary 11.17. *The subgroups of $\mathbb{Z}$ are exactly $n\mathbb{Z}$ for $n = 0, 1, 2, \ldots$.*

Order

Definition 11.18 (Order). Let G be a group, $g \in G$. If there is a positive integer k such that $a^k = 1$, then the ***order*** of g is

$$|g| := \min\{m > 0 \mid a^m = 1\}.$$

Otherwise we say that the order of g is infinite.

We have given two different meanings to the word "order". One is the order of a group and the other is the order of an element. Since mathematicians are usually (but not always) sensible, the name wouldn't be used twice if they weren't related. This is explained by the next result.

Lemma 11.19. *For $g \in G$, $|g| = |\langle g \rangle|$.*

Proof. We consider the cases where $|g|$ is finite or infinite.

Case 1: $|g| = \infty$**.** Then $g^n \neq g^m$ for all $n \neq m$; otherwise $g^{m-n} = e$, which contradicts the minimality of $|g|$. Hence $|\langle g \rangle| = \infty = |g|$.

Case 2: $|g| < \infty$**.** Suppose $|g| = k$. Then $g^k = 1$.

Claim. $\langle g \rangle = \{e, g, g^2, \ldots, g^{k-1}\}$.

Note that $\langle g \rangle$ does not contain higher powers of g, since $g^k = 1$ so higher powers will loop back to existing elements. There are also no repeating elements in the list provided since $g^m = g^n$ implies $g^{m-n} = e$. Hence $|\langle g \rangle| = k = |g|$.

□

Lemma 11.20. *If $g \in G$ and $|g|$ is finite, then $g^n = e$ if and only if $|g| \mid n$.*

Proof.

$\boxed{\Longleftarrow}$ Suppose $|g| \mid n$. Then $n = k|g|$ for some $k \in \mathbb{Z}$, so

$$g^n = (g^{|g|})^k = e^k = e.$$

$\boxed{\Longrightarrow}$ Suppose $g^n = e$. By the division algorithm, there exists integers q, r such that $n = q|g| + r$, where $0 \leq r < |g|$. Then

$$g^r = g^{n-q|g|} = g^n (g^{|g|})^{-q} = e.$$

By the minimality of $|g|$, we must have $r = 0$, and so $n = q|g|$ implies $|g| \mid n$. □

Corollary 11.21. *Let G be a cyclic group, $g \in G$. Then $g^m = g^n \iff m \equiv n \pmod{|g|}$.*

Proof. $g^m = g^n \iff g^{m-n} = e \iff |g| \mid m-n \iff m \equiv n \pmod{|g|}$. □

11.2 Homomorphisms and Isomorphisms

In this section, we make precise the notion of when two groups "look the same"; that is, they have the same group-theoretic structure. This is the notion of an *isomorphism* between two groups.

Definitions and Properties

When we talk about functions between groups it makes sense to limit our scope to functions that preserve the group operation (morphisms in the category of groups). More precisely:

Definition 11.22 (Homomorphism)**.** Let $(G,*)$ and $(H,\diamond)$ be groups. We say $\phi: G \to H$ is a ***homomorphism*** if

$$\phi(x*y) = \phi(x)\diamond\phi(y) \quad (x,y \in G).$$

When the group operations for G and H are understood, we omit them and simply write

$$\phi(xy) = \phi(x)\phi(y).$$

Example.

- Let G be a commutative group. The map $x \mapsto x^{-1}$ from G into itself is a homomorphism.
- The map $z \mapsto |z|$ is a homomorphism from $\mathbb{C}^\times$ to $\mathbb{R}^+$.
- The map $x \mapsto e^x$ is a homomorphism from $(\mathbb{R},+)$ to $(\mathbb{R}^+,\times)$. Its inverse map, the logarithm, is also a homomorphism.

Lemma 11.23 (Basic properties)**.** *Let $\phi: G \to H$ be a homomorphism. Let $g \in G$, $n \in \mathbb{Z}$. Then*

(i) $\phi(e_G) = e_H$

(ii) $\phi(g^{-1}) = \phi(g)^{-1}$

(iii) $\phi(g^n) = \phi(g)^n$

Proof.

(i) $\phi(e_G) = \phi(e_G e_G) = \phi(e_G)\phi(e_G)$, then apply $\phi(e_G)^{-1}$ to both sides to get $\phi(e_G) = e_H$.

(ii) $\phi(g)\phi(g^{-1}) = \phi(gg^{-1}) = \phi(e_G) = e_H$.

(iii) Note more generally that we can show $\phi(g^n) = (\phi(g))^n$ for $n > 0$ by induction. For $n = -k < 0$ we have

$$\phi(g^n) = \phi((g^{-1})^k) = \phi(g^{-1})^k = (\phi(g)^{-1})^k = \phi(g)^n.$$

□

Lemma 11.24. *The composition of homomorphisms is a homomorphism.*

Proof. Suppose $\phi : G \to H$ and $\psi : H \to K$ are homomorphisms. We have

$$(\psi \circ \phi)(xy) = \psi(\phi(xy)) = \psi(\phi(x)\phi(y)) = \psi(\phi(x))\psi(\phi(y)) = (\psi \circ \phi)(x)(\psi \circ \phi)(y).$$

Hence $\psi \circ \phi$ is a homomorphism. □

Let $\mathrm{Hom}(G,H)$ denote the set of homomorphisms from G to H. Then $\mathrm{Hom}(G,H)$ is a group under addition.

(i) If $\phi, \psi \in \mathrm{Hom}(G,H)$, then for $x, y \in G$,

$$\begin{aligned}(\phi+\psi)(x+y) &= \phi(x+y) + \psi(x+y) \\ &= \phi(x) + \psi(x) + \phi(y) + \psi(y) \\ &= (\phi+\psi)(x) + (\phi+\psi)(y),\end{aligned}$$

so that $f+g$ is a homomorphism.

(ii) If $\phi, \psi, \gamma \in \mathrm{Hom}(G,H)$, then for all $x \in G$,

$$((\phi+\psi)+\gamma)(x) = (\phi+\psi)(x) + \gamma(x) = \phi(x) + \psi(x) + \gamma(x),$$

and

$$(\phi+(\psi+\gamma))(x) = \phi(x) + (\psi+\gamma)(x) = \phi(x) + \phi(x) + \gamma(x).$$

Hence $(\phi+\psi)+\gamma = \phi+(\psi+\gamma)$.

(iii) The zero map is the identity element of $\mathrm{Hom}(G,H)$.

(iv) The inverse of $\phi \in \mathrm{Hom}(G,H)$ is $-\phi$ (which is a homomorphism).

Definition 11.25 (Isomorphism)**.** An ***isomorphism*** is a bijective homomorphism. If there exists an isomorphism $\phi : G \to H$, we say G and H are ***isomorphic***, denoted by $G \cong H$.

An *automorphism* of a group G is an isomorphism from G to G; the automorphisms of G form a group $\mathrm{Aut}(G)$ under composition. An *endomorphism* of G is a homomorphism from G to G.

Example. The exponential map $\exp\colon \mathbb{R} \to \mathbb{R}^+$ defined by $\exp(x) = e^x$ is an isomorphism from $(\mathbb{R}, +)$ to $(\mathbb{R}^+, \times)$.

(i) exp is a bijection since it has an inverse function (namely ln).

(ii) exp preserves the group operations since $e^{x+y} = e^x e^y$.

Hence $(\mathbb{R}, +) \cong (\mathbb{R}^+, \times)$.

Lemma 11.26. *G is abelian if and only if $\phi\colon G \to G$ defined by $\phi(x) = x^{-1}$ is an isomorphism.*

Proof.

$\boxed{\Longrightarrow}$ Suppose G is abelian. Then

$$\phi(xy) = (xy)^{-1} = y^{-1}x^{-1} = x^{-1}y^{-1} = \phi(x)\phi(y)$$

so ϕ is a homomorphism.

$\boxed{\Longleftarrow}$ Suppose $\phi(x) = x^{-1}$ is an isomorphism. Then

$$ab = \phi(a^{-1})\phi(b^{-1}) = \phi(a^{-1}b^{-1}) = (a^{-1}b^{-1})^{-1} = ba.$$

□

If $G \cong H$, then any property of G that depends only on the group structure will also hold for H:

Lemma 11.27. *Suppose $G \cong H$.*

(i) $|G| = |H|$.

(ii) *G is abelian if and only if H is abelian.*

(iii) *For all $g \in G$, $|g| = |\phi(g)|$.*

(iv) $G' \leq G \iff \phi(G') \leq H$.

Proof. Suppose $\phi\colon G \to H$ is an isomorphism.

(i) This follows since ϕ is a bijection.

(ii) Suppose G is abelian. Then for all $a, b \in G$, $ab = ba$. This implies $\phi(ab) = \phi(ba)$, so $\phi(a)\phi(b) = \phi(b)\phi(a)$. Hence H is abelian.

Now suppose H is abelian. Then for all $a, b \in H$, $ab = ba$. Since ϕ is an isomorphism, ϕ^{-1} is an isomorphism. Then $\phi^{-1}(ab) = \phi^{-1}(ba)$, so $\phi^{-1}(a)\phi^{-1}(b) = \phi^{-1}(b)\phi^{-1}(a)$. Hence G is abelian.

(iii) Let $g \in G$, $|g| = n$. Then $g^n = e$, so $\phi(g^n) = \phi(g)^n = e$. Thus $|\phi(g)| \leq n$.

Similarly, suppose $|\phi(g)| = n$. Then $\phi(g)^n = \phi(g^n) = e = \phi(e)$, so $g^n = e$ by injectivity. Thus $|g| \leq n$.

Hence $|g| = |\phi(g)|$.

(iv)

□

Example. Consider the quaternion group Q_8. Note that $|\pm i| = |\pm j| = |\pm k| = 4$. In particular, no element has order 8. Hence $Q_8 \ncong \mathbb{Z}/8\mathbb{Z}$.

Kernel and Image

We introduce two important groups related to every homomorphism.

Definition 11.28 (Kernel). Let $\phi : G \to H$ be a homomorphism. The ***kernel*** of ϕ is

$$\ker \phi := \{g \in G \mid \phi(g) = e_H\}.$$

Lemma. $\ker \phi \triangleleft G$.

Proof. Apply the subgroup criterion. Since $e_G \in \ker \phi$, $\ker \phi \neq \emptyset$. Let $x, y \in \ker \phi$; that is, $\phi(x) = \phi(y) = e_H$. Then

$$\phi(xy^{-1}) = \phi(x)\phi(y)^{-1} = e_H$$

so $xy^{-1} \in \ker \phi$. By the subgroup criterion, $\ker \phi \leq G$.

Let $x \in \ker \phi$, $g \in G$. Then

$$\phi(gxg^{-1}) = \phi(g)\phi(x)\phi(g^{-1}) = 1,$$

so $gxg^{-1} \in \ker \phi$. Hence $\ker \phi \triangleleft G$. □

Definition 11.29 (Image). Let $\phi : G \to H$ be a homomorphism. The ***image*** of G under ϕ is

$$\operatorname{im} \phi := \phi(G) = \{\phi(g) \mid g \in G\}.$$

Remark. $\operatorname{im} \phi$ is the usual set theoretic image of ϕ.

Lemma. $\operatorname{im}\phi \leq H$.

Proof. Since $\phi(e_G) = e_H$, $e_H \in \operatorname{im}\phi$ so $\operatorname{im}\phi \neq \emptyset$. Let $x, y \in \operatorname{im}\phi$. Then there exists $a, b \in G$ such that $\phi(a) = x$, $\phi(b) = y$. Then

$$xy^{-1} = \phi(a)\phi(b)^{-1} = \phi(ab^{-1})$$

so $xy^{-1} \in \operatorname{im}\phi$. By the subgroup criterion, $\operatorname{im}\phi \leq G$. □

The following result is a useful characterisation for injective homomorphisms.

Lemma 11.30. *Let $\phi : G \to H$ be a homomorphism. Then ϕ is injective if and only if* $\ker\phi = \{e_G\}$.

Proof.

$\boxed{\Longrightarrow}$ Suppose ϕ is injective. Since $\phi(e_G) = e_H$, $e_G \in \ker\phi$ so $\{e_G\} \subset \ker\phi$.

Conversely, let $x \in \ker\phi$, so $\phi(x) = e_H$. Then $\phi(x) = e_H = \phi(e_G)$, so by injectivity $x = e_G$. Hence $\ker\phi \subset \{e_G\}$, so $\ker\phi = \{e_G\}$.

$\boxed{\Longleftarrow}$ Suppose $\ker\phi = \{e_G\}$. Suppose $\phi(a) = \phi(b)$, then $\phi(ab^{-1}) = \phi(a)\phi(b^{-1}) = \phi(a)\phi(a)^{-1} = e_H$. Hence $ab^{-1} \in \ker\phi = \{e_G\}$, so $ab^{-1} = e_G$ and thus $a = b$. Therefore ϕ is injective. □

Lemma 11.31. *Let $\phi : G \to H$ be an isomorphism. Then its inverse $\phi^{-1} : H \to G$ is an isomorphism.*

Proof. The inverse of a bijective map is bijective. Hence it suffices to show that $\phi^{-1}(x)\phi^{-1}(y) = \phi^{-1}(xy)$ for all $x, y \in H$.

Let $a = \phi^{-1}(x)$, $b = \phi^{-1}(y)$, $c = \phi^{-1}(xy)$; we will show that $ab = c$. Since ϕ is bijective, it suffices to show that $\phi(ab) = \phi(c)$.

Since ϕ is a homomorphism,

$$\phi(ab) = \phi(a)\phi(b) = xy = \phi(c).$$

□

Cosets

Definition 11.32 (Coset). Let $H \leq G$. For $a \in G$, a ***left coset*** and ***right coset*** of H in G

are

$$aH := \{ah \mid h \in H\}$$
$$Ha := \{ha \mid h \in H\}$$

Any element of a coset is called a *representative* for the coset.

Example. Consider the subgroup $2\mathbb{Z} \leq \mathbb{Z}$. Then $6+2\mathbb{Z} = \{\text{all even numbers}\} = 0+2\mathbb{Z}$, and $1+2\mathbb{Z} = \{\text{all odd numbers}\} = 17+2\mathbb{Z}$.

Notation. We denote the set of (left) cosets by G/H.

In what will follow, the analogous results hold similarly for right cosets.

Lemma 11.33. *Let $H \leq G$. Then $aH = H$ if and only if $a \in H$.*

Proof.

$\boxed{\Longrightarrow}$ Suppose $aH = H$. Then $ah \in H$ for some $h \in H$. Let $k = ah$, then $a = kh^{-1} \in H$.

$\boxed{\Longleftarrow}$ Let $a \in H$. Then $aH \subset H$.

Since $a^{-1} \in H$, $a^{-1}H \subset H$. Then $H = eH = (aa^{-1})H = a(a^{-1})H \subset aH$. Hence $aH = H$. □

The next result shows when two cosets are equal.

Lemma 11.34. *Let $H \leq G$, $a, b \in G$. Then $aH = bH$ if and only if $a^{-1}b \in H$.*

Proof.

$$\begin{aligned} aH = bH &\iff a^{-1}(aH) = a^{-1}bH \\ &\iff (a^{-1}a)H = (a^{-1}b)H \\ &\iff H = (a^{-1}b)H \end{aligned}$$

From the previous result, $H = (a^{-1}b)H$ if and only if $a^{-1}b \in H$. □

Proposition 11.35. *Let $H \leq G$. Then G/H forms a partition of G.*

We need to prove the following.

(i) For all $a \in G$, $aH \neq \emptyset$.

(ii) $\bigcup_{a \in G} aH = G$.

(iii) For every $a, b \in G$, $aH \cap bH = \emptyset$ or $aH = bH$.

Proof.

(i) Since $H \leq G$, $1 \in H$. Thus for all $a \in G$, $a = a1 \in aH$ so $aH \neq \emptyset$.

(ii) For all $a \in G$, $aH \subset G$, then $\bigcup_{a\in G} aH \subset G$. Note that $a \in G$ implies $a = ae \in aH$, and so $G = \bigcup_{a\in G} g \subset \bigcup_{a\in G} aH$. By double inclusion we are done.

(iii) If $aH \cap bH = \emptyset$, then we are done. If $aH \cap bH \neq \emptyset$ we need to show $aH = bH$. Let $x \in G$ such that $x \in aH \cap bH$. Then $x = ah_1 = bh_2$ for $h_1, h_2 \in H$ so $h_1 = a^{-1}bh_2$. Notice that $a^{-1}b = h_1h_2^{-1} \in H$ and thus $aH = bH$.

□

The next result shows that the left cosets of H partition G into equal-sized parts.

Lemma 11.36. *The cosets of H in G are the same size as H; that is, for all $a \in G$, $|aH| = |H|$.*

Proof. Consider the mapping

$$\begin{aligned} f : H &\to aH \\ h &\mapsto ah \end{aligned}$$

We will show that f is bijective.

- Let $h_1, h_2 \in H$, then

$$\begin{aligned} f(h_1) = f(h_2) &\implies ah_1 = ah_2 \\ &\implies a^{-1}ah_1 = a^{-1}ah_2 \\ &\implies h_1 = h_2 \end{aligned}$$

  so f is injective.

- Note that f is surjective by the definition of aH.

Since f is bijective, $|H| = |aH|$. □

Lagrange's Theorem

Definition 11.37 (Index). Let $H \leq G$. The ***index*** of H in G is the number of left cosets of H in G, denoted by $|G : H|$.

Then $|G| = |G : 1|$; that is, the order of G is the index of the trivial subgroup in G.

Theorem 11.38 (Lagrange's theorem). *Let G be a finite group, $H \leq G$. Then $|H|$ divides $|G|$; in particular,*

$$|G| = |H|\,|G:H|. \tag{11.1}$$

(11.1) is known as the *counting formula*.

Proof. Suppose that there are $|G:H|$ left cosets in total. Since the left cosets partition G, and each coset has size $|H|$, we have

$$|H|\,|G:H| = |G|.$$

□

Corollary 11.39. *The order of an element of a finite group divides the order of the group.*

Proof. Consider the subgroup generated by a, which has order $o(a)$. Then by Lagrange's theorem, $o(a)$ divides $|G|$. □

Corollary 11.40. *For any finite group G and $a \in G$, $a^{|G|} = 1$.*

Proof. We know that $|G| = k\,o(a)$ for some $k \in \mathbb{N}$. Then $a^{|G|} = \left(a^{o(a)}\right)^k = 1^k = 1$. □

The above result has a well-known special case in modular arithmetic.

Corollary 11.41 (Fermat's little theorem). *If p is prime, then $a^p \equiv a \pmod{p}$ for any $a \in \mathbb{Z}$.*

Proof. Take $G = (\mathbb{Z}/p\mathbb{Z})^\times = \{1, \ldots, p-1\}$. □

Corollary 11.42. *A group of prime order is cyclic.*

Proof. Let $|G| = p$ be prime. Let $a \in G$, $a \neq 1$. We will show that $G = \langle a \rangle$.

Since $o(a) \mid |G| = p$ and $o(a) > 1$, we must have $o(a) = p$. Notice that this is also the order of $\langle a \rangle$. Since G has order p, thus $\langle a \rangle = G$. □

This corollary classifies groups of prime order p. They form one isomorphism class: the class of the cyclic groups of order p.

The next result is of great interest in number theory. The *Euler ϕ-function* $\phi(n)$ is defined for all positive integers as follows:

$$\phi(n) = \begin{cases} 1 & (n = 1) \\ \text{number of positive integers less than } n\text{, relatively prime to } n & (n > 1) \end{cases}$$

Theorem 11.43 (Euler). *If n is a positive integer, and a is coprime to n, then*

$$a^{\phi(n)} \equiv 1 \pmod{n}.$$

Counting Principle

We generalise the notion of cosets, as defined earlier.

Definition 11.44. Let $H, K \leq G$. Define

$$HK := \{hk \mid h \in H, k \in K\}.$$

Lemma 11.45. *Let $H, K \leq G$. Then $HK \leq G$ if and only if $HK = KH$.*

Proof.

$\boxed{\Longleftarrow}$ Suppose $HK = KH$; that is, if $h \in H$ and $k \in K$, then $hk = k_1 h_1$ for some $k_1 \in K, h_1 \in H$. We now show that HK is a subgroup of G:

(i) $1 \in H$ and $1 \in K$, so $1 \in HK$.

(ii) Let $x = hk \in HK$, $y = h'k' \in HK$. then

$$xy = hkh'k'.$$

Note that $kh' \in KH = HK$, so $kh' = h_2 k_2$ for some $h_2 \in H, k_2 \in K$. Then

$$xy = h(h_2 k_2)k' = (hh_2)(k_2 k') \in HK.$$

Thus HK is closed.

(iii) Let $x \in HK$, then $x = hk$ for some $h \in H, k \in K$. Thus

$$x^{-1} = (hk)^{-1} = k^{-1}h^{-1} \in KH = HK,$$

so $x^{-1} \in HK$.

$\boxed{\Longrightarrow}$ Suppose $HK \leq G$.

- Let $x \in KH$, so $x = kh$ for some $k \in K, h \in H$. Then

$$x = kh = (h^{-1}k^{-1})^{-1} \in HK.$$

Thus $KH \subset HK$.

- Let $x \in HK$. Since $HK \leq G$, HK is closed under inverses, so $x^{-1} = hk \in HK$. Then

$$x = (x^{-1})^{-1} = (hk)^{-1} = k^{-1}h^{-1} \in KH.$$

  Thus $HK \subset KH$.

Hence $HK = KH$. □

An interesting special case is the situation when G is an abelian group, for in that case trivially $HK = KH$. Thus as a consequence we have the following result.

Corollary 11.46. *Let $H, K \leq G$, where G is abelian. Then $HK \leq G$.*

Proposition 11.47. *If $H, K \leq G$ are finite groups, then*

$$|HK| = \frac{|H||K|}{|H \cap K|}.$$

Proof. Notice that HK is a union of left cosets of K, namely

$$HK = \bigcup_{h \in H} hK.$$

□

Normal Subgroups, Quotient Groups

Definition 11.48 (Normal subgroup)**.** Let G be a group. We say $H \leq G$ is a ***normal subgroup*** of G, denoted by $H \triangleleft G$, if

$$aH = Ha \quad (\forall a \in G)$$

Remark. This does *not* mean that $ah = ha$ for all $a \in G$, $h \in H$ or that G is abelian; although we can easily see that all subgroups of abelian groups are normal. In general, a left coset does not equal the right coset.

Lemma 11.49. *The following are equivalent.*

(i) $H \triangleleft G$.

(ii) $ghg^{-1} \in H$ *for all* $g \in G$, $h \in H$.

(iii) $gHg^{-1} = H$ *for all* $g \in G$.

Proof.

(i) $\iff$ (ii) First suppose $aH = Ha$ for all $a \in G$. Let $g \in G$, $x \in H$. Then $gH = Hg$ so $gx = h'g$ for some $h' \in H$. Then $gxg^{-1} = h'gg^{-1} = h' \in H$.

Conversely suppose $ghg^{-1} \in H$ for all $g \in G$, $h \in H$. Fix g. Then $ghg^{-1} \in H$ implies $gh \in Hg$ for all $h \in H$. So $gH \subset Hg$. Similarly $gH \supset Hg$, so $gH = Hg$.

(i) $\iff$ (iii) $H \triangleleft G$ if and only if for all $g \in G$,

$$\begin{aligned} gH = Hg &\iff (gH)g^{-1} = (Hg)g^{-1} \\ &\iff gHg^{-1} = H \end{aligned}$$

□

Remark. We frequently use (ii) to check if a subgroup is a normal subgroup.

Lemma 11.50. *A subgroup of an abelian group is normal.*

Proof. Let G be abelian, $H \leq G$. For all $g \in G$, $h \in H$, we have $ghg^{-1} = gg^{-1}h = h \in H$. Thus H is normal. □

Lemma 11.51. *The intersection of normal subgroups is a normal subgroup.*

Proof. Suppose H_1 and H_2 are normal. Then for all $g \in G$,

- $gH_1g^{-1} \subset H_1$,
- $gH_2g^{-1} \subset H_2$.

Since $H_1 \cap H_2$ is a subset of both H_1 and H_2, we must have that for all $g \in G$,

- $g(H_1 \cap H_2)g^{-1} \subset H_1$,
- $g(H_1 \cap H_2)g^{-1} \subset H_2$.

Together these imply that for all $g \in G$,

$$g(H_1 \cap H_2)g^{-1} \subset H_1 \cap H_2.$$

Hence $H_1 \cap H_2$ is normal. □

The (left) cosets of a group form a group, known as the *quotient group*.

Definition 11.52 (Quotient group). Let G be a group, $H \triangleleft G$. Then the ***quotient group***

of G by H is the set of left cosets of H in G:

$$G/H := \{aH \mid a \in G\}.$$

Remark. Quotient groups are not subgroups of G; they contain different kinds of elements. For example, $\mathbb{Z}/n\mathbb{Z} \cong C_n$ are finite, but all subgroups of $\mathbb{Z}$ infinite.

Lemma. *G/H is a group under the operation $aH * bH = (ab)H$.*

Proof. First show that the operation is well-defined; that is, if $aH = a'H$ and $bH = b'H$, we want to show that $aH * bH = a'H * b'H$.

We know that $a' = ak_1$ and $b' = bk_2$ for some $k_1, k_2 \in H$. Then $a'b' = ak_1bk_2$. We know that $b^{-1}k_1b \in H$. Let $b^{-1}k_1b = k_3$. Then $k_1b = bk_3$. So $a'b' = abk_3k_2 \in (ab)H$. So picking a different representative of the coset gives the same product.

If aH and bH are cosets, then $(ab)H$ is also a coset, so the operation is closed.

(i) For $a, b, c \in G$, by associativity of G,

$$(aH)(bHcH) = (aH)(bcH) = a(bc)H = (ab)cH = (aHbH)cH$$

so the operation is associative.

(ii) The identity is $1H = \{1h \mid h \in H\} = \{h \mid h \in H\} = H$.

(iii) The inverse of aH is $a^{-1}H$, since

$$(aH)(a^{-1}H) = aa^{-1}H = H \implies (aH)^{-1} = a^{-1}H.$$

□

Example (Modular arithmetic)**.** Fix $n \in \mathbb{Z}^+$. Evidently $n\mathbb{Z}$ is a subgroup of $\mathbb{Z}$. Then the quotient group $\mathbb{Z}/n\mathbb{Z}$ consists of cosets of the form

$$n\mathbb{Z}, 1 + n\mathbb{Z}, 2 + n\mathbb{Z}, \ldots, n-1 + n\mathbb{Z}.$$

If we consider each coset as an equivalence class, we write

$$\mathbb{Z}/n\mathbb{Z} = \{[0], [1], \ldots, [n-1]\}.$$

Addition on $\mathbb{Z}/n\mathbb{Z}$ is defined as

$$[x] + [y] = [x + y].$$

The next result concerns the order of the quotient group.

Lemma 11.53. *Let G be a finite group, $H \triangleleft G$. Then*

$$|G/H| = |G : H| = \frac{|G|}{|H|}.$$

Proof. Since G/H has as its elements the left cosets of H in G, and there are precisely $|G : H|$ such cosets, the first equality holds.

The second equality holds by Lagrange's theorem. □

We now define a *canonical* homomorphism ("natural" map) from a group to its quotient group.

Definition 11.54 (Quotient map)**.** Let $H \triangleleft G$. The ***quotient map*** is the map

$$\begin{aligned} \pi \colon G &\to G/H \\ a &\mapsto aH \end{aligned}$$

Lemma 11.55. *Quotient maps are surjective homomorphisms.*

Proof. Let $\pi \colon G \to G/H$ which maps $a \mapsto aH$ be a quotient map.

- For all $a, b \in G$,
$$\pi(ab) = (ab)H = (aH)(bH) = \pi(a)\pi(b).$$
Thus π is a homomorphism.
- For all $aH \in G/H$, $\pi(a) = aK$. Thus π is surjective.

□

The next result provides a characterisation of normal subgroups.

Lemma 11.56. *$H \triangleleft G$ if and only if H is the kernel of some homomorphism.*

Proof.

$\boxed{\Longleftarrow}$ Suppose $H = \ker \phi$ for some homomorphism $\phi \colon G \to G'$.

Let $g \in G$, $h \in H$. Then

$$\phi(ghg^{-1}) = \phi(g)\phi(h)\phi(g)^{-1} = \phi(h) = 1.$$

Thus $ghg^{-1} \in \ker \phi = H$.

$\boxed{\Longrightarrow}$ The kernel of the quotient map is H itself:

$$\ker \pi = \{a \in G \mid aH = H\} = \{a \in G \mid a \in H\} = H.$$

□

Isomorphism Theorems

In this section, we will prove several isomorphism theorems.

Theorem 11.57 (First isomorphism theorem)**.** *Let $\phi\colon G \to H$ be a homomorphism. Then*

$$G/\ker\phi \cong \operatorname{im}\phi. \tag{11.2}$$

Proof. For ease of notation, denote $K = \ker\phi$. Consider the map $\theta\colon G/K \to \operatorname{im}\phi$. We claim
$$\begin{aligned} G/K &\to \operatorname{im}\phi \\ xK &\mapsto \phi(x) \end{aligned}$$
that θ is an isomorphism.

1. We check that θ is well-defined. Let $x, y \in G$, suppose $xK = yK$. Then

$$\begin{aligned} & xK = yK \\ \iff & x^{-1}y \in K \\ \iff & \phi(x^{-1}y) = e_H \\ \iff & \phi(x)^{-1}\phi(y) = e_H \\ \iff & \phi(x) = \phi(y) \end{aligned}$$

2. θ is a homomorphism: $\theta(xKyK) = \theta(xyK) = \phi(xy) = \phi(x)\phi(y) = \theta(xK)\theta(yK)$.

3. θ is bijective:

   - θ is injective since $\theta(xK) = \theta(yK) \implies \phi(x) = \phi(y) \implies xK = yK$.
   - θ is surjective, since $\operatorname{im}\theta = \{\theta(xK) \mid x \in G\} = \{\phi(x) \mid x \in G\} = \operatorname{im}\phi$.

□

Corollary 11.58. *Any cyclic group is isomorphic to either $\mathbb{Z}$ or $\mathbb{Z}/n\mathbb{Z}$ for some $n \in \mathbb{N}$.*

Proof. Let $G = \langle g \rangle$ for some $g \in G$. Define the map $\phi\colon \mathbb{Z} \to G$. We claim that ϕ is a surjective
$$\begin{aligned} \mathbb{Z} &\to G \\ m &\mapsto g^m \end{aligned}$$
homomorphism.

1. ϕ is a homomorphism, since $\phi(m_1 + m_2) = g^{m_1+m_2} = g^{m_1}g^{m_2} = \phi(m_1)\phi(m_2)$.

2. ϕ is surjective, since G is by definition all g^m for all m.

By surjectivity, $\operatorname{im}\phi = G$. We know that $\ker\phi \triangleleft \mathbb{Z}$. We have the following possibilities for the kernel:

Case 1: $\ker\phi = \{1\}$ This implies ϕ is injective, so ϕ is an isomorphism. Hence $G \cong \mathbb{Z}$.

Case 2: $\ker\phi = \mathbb{Z}$ By the first isomorphism theorem, $G \cong \mathbb{Z}/\mathbb{Z} = \{1\} = C_1$.

Case 3: $\ker\phi = n\mathbb{Z}$ (Since these are the only remaining proper subgroups of $\mathbb{Z}$.) By the first isomorphism theorem, $G \cong \mathbb{Z}/n\mathbb{Z}$.

□

Example (Circle group)**.** Consider the subgroup $(\mathbb{Z},+)$ of $(\mathbb{R},+)$. The quotient group $\mathbb{R}/\mathbb{Z}$ is called the *circle group*.
Define a congruence relation on $\mathbb{R}$:

$$x \sim y \iff x - y \in \mathbb{Z}.$$

If $x \sim y$, we say $x, y \in \mathbb{R}$ are *congruent* mod $\mathbb{Z}$, and denote $x \equiv y \pmod{\mathbb{Z}}$. This congruence is an equivalence relation, and the congruence classes are precisely the cosets of $\mathbb{Z}$ in $\mathbb{R}$.
If $x \equiv y \pmod{\mathbb{Z}}$, then $e^{2\pi ix} = e^{2\pi iy}$, and conversely. Thus the map $f\colon \mathbb{R}/\mathbb{Z} \to \mathbb{T}$ is $x \mapsto e^{2\pi ix}$ an isomorphism, where $\mathbb{T} = \{z \in \mathbb{C} \mid |z| = 1\}$ is the multiplicative group of complex numbers having absolute value 1.

Remark. $2\pi x$ can be considered as the angle measured from the positive real axis of $\mathbb{C}$.

Example. Let $\mathbb{C}^\times$ be the multiplicative group of non-zero complex numbers, and $\mathbb{R}^+$ the multiplicative group of positive real numbers. Given a complex number $z \neq 0$, we can write

$$z = ru,$$

where $r \in \mathbb{R}^+$, and u has absolute value 1. (Let $u = z/|z|$.) Such an expression is uniquely determined, and the map

$$\begin{aligned} f\colon \mathbb{C}^\times &\to \mathbb{T} \\ z &\mapsto \frac{z}{|z|} \end{aligned}$$

is a homomorphism. Since $\ker f = \mathbb{R}^+$ and $\operatorname{im} f = \mathbb{T}$ (by surjectivity), by the first isomorphism theorem, $\mathbb{C}^\times/\mathbb{R}^+$ is isomorphic to $\mathbb{T}$.

Theorem 11.59 (Second isomorphism theorem)**.** *Let $H \leq G$, $K \triangleleft G$. Then*

$$HK/K \cong H/(H \cap K). \tag{11.3}$$

We first prove a few preliminary results.

Lemma. *Let $H \leq G$, $K \triangleleft G$. Then*

(i) $HK \leq G$;

(ii) $K \triangleleft HK$;

(iii) $H \cap K \triangleleft H$.

Proof.

(i) Since $1 \in H$ and $1 \in K$, we have $1 \in HK$, so $HK \neq \emptyset$. Let $hk, h'k' \in HK$. Then

$$h'k'(hk)^{-1} = h'k'k^{-1}h^{-1} = (\underbrace{h'h^{-1}}_{\in H})(\underbrace{hk'k^{-1}h^{-1}}_{\in K, \text{ by normality}}) \in HK.$$

By the subgroup criterion, $HK \leq G$.

(ii)

(iii) Since the intersection of subgroups is a subgroup, $H \cap K$ is a subgroup of N. It remains to be shown that $H \cap K$ is normal in H.

Let $h \in H$, $x \in H \cap K$. We will show that $hxh^{-1} \in H \cap K$.

$H \leq G$ and $h \in H$ imply $h \in G$. Since $K \triangleleft G$, $x \in H \cap K$ and $h \in H$ imply $hxh^{-1} \in H$.

□

We are now ready to prove the theorem.

Proof. Define the map $\phi : H \to G/K$.
$$h \mapsto hK$$
ϕ is a homomorphism: $\phi(xy) = (xy)K = (xK)(yK) = \phi(x)\phi(y)$.
The kernel and image of ϕ are

$$\ker\phi = \{h \in H \mid hK = K\} = \{h \in H \mid h \in K\} = H \cap K,$$
$$\operatorname{im}\phi = \{\phi(h) \mid h \in H\} = \{hK \mid h \in H\} = HK/K.$$

Hence the desired result follows from the first isomorphism theorem. □

Theorem 11.60 (Third isomorphism theorem)**.** *Let $H, K \triangleleft G$, $H \leq K$. Then*

$$(G/H)/(K/H) \cong G/K. \tag{11.4}$$

Lemma. $K/H \triangleleft G/H$.

Proof. We first show $K/H \leq G/H$:

(i) The identity of G/H is H, which is also the identity of K/H, since $1 \in K$.

(ii) Let $aH, bH \in K/H$. Since $ab \in K$ for all $a, b \in K$, we have $(aH)(bH) = (ab)H \in K/H$.

(iii) Let $aH \in K/H$. Since $a^{-1} \in K$, we have $a^{-1}H \in K/H$.

To show normality, let $gH \in G/H$, $kH \in K/H$. Then $(gH)(kH)(gH)^{-1} = (gkg^{-1})H$. Since $K \triangleleft G$, $gkg^{-1} \in K$. Thus $(gkg^{-1})H \in K/H$. □

We can now prove the theorem.

Proof. Define the *canonical* homomorphism $\phi : G/H \to G/K$.

$$\begin{aligned} gH &\mapsto gK \end{aligned}$$

We claim that ϕ is a surjective homomorphism.

1. We check that ϕ is well-defined: If $gH = g'H$, then $g^{-1}g' \in H$. Since $H \subset K$, $g^{-1}g' \in K$. Thus $gK = g'K$.
2. ϕ is a homomorphism: $\phi(gHg'H) = \phi(gg'H) = gg'K = (gK)(g'K) = \phi(gH)\phi(g'H)$.
3. ϕ is clearly surjective, since any coset gK is the image $\phi(gH)$.

The kernel and image of ϕ are

$$\begin{aligned} \ker\phi &= \{gH \mid gK = K\} = \{gH \mid g \in K\} = K/H, \\ \operatorname{im}\phi &= G/K \quad \text{by surjectivity.} \end{aligned}$$

Hence the conclusion follows from the first isomorphism theorem. □

We now discuss an isomorphism theorem concerning pre-images of groups.

Theorem 11.61. *Let $\phi : G \to G'$ be a surjective homomorphism. Let $H' \triangleleft G'$ and $H = \phi^{-1}(H')$. Then*

$$G/H \cong G'/H'. \tag{11.5}$$

Lemma. $H = \phi^{-1}(H') \triangleleft G$.

Proof. We first show $H \leq G$.

(i) Since $H' \leq G'$, $e_{G'} \in H'$. Then $\phi(e_G) = e_{G'} \in H'$, so $e_G \in \phi^{-1}(H')$.

(ii) Let $a, b \in H$. Then $\phi(a), \phi(b) \in H'$. By closure, $\phi(a)\phi(b) = \phi(ab) \in H'$, so $ab \in \phi^{-1}(H')$.

(iii) Let $a \in H$. Then $\phi(a) \in H'$. Since H' is closed under inverses, $\phi(a)^{-1} = \phi(a^{-1}) \in H'$, so $a^{-1} \in \phi^{-1}(H')$.

To show normality, let $g \in G$, $h \in H$, then $\phi(h) \in H'$. Since $\phi(ghg^{-1}) = \phi(g)\phi(h)\phi(g)^{-1}$, where $\phi(g) \in G'$ and $\phi(h) \in H'$, by normality $\phi(ghg^{-1}) \in H'$. Thus $ghg^{-1} \in \phi^{-1}(H')$. □

We now prove the theorem.

Proof. Consider the map $\psi\colon G \to G'/H'$. Note that ψ can also be described as the composite map $\psi = \pi \circ \phi$:

$$\begin{aligned} g &\mapsto \phi(g)H' \\ G \xrightarrow{\phi} G' &\xrightarrow{\pi} G'/H' \end{aligned}$$

where $\pi\colon G' \to G'/H'$ is the quotient map.

We claim that ψ is a surjective homomorphism.

1. The composition of homomorphisms is a homomorphism, so ψ is a homomorphism.
2. The composition of surjective maps is surjective, so ψ is surjective.

The kernel and image of ψ are

$$\begin{aligned} \ker\psi &= \{g \in G \mid \psi(g) = H'\} = \{g \in G \mid \phi(g)H' = H'\} \\ &= \{g \in G \mid \phi(g) \in H'\} = \phi^{-1}(H') = H, \\ \operatorname{im}\psi &= G'/H' \quad \text{by surjectivity.} \end{aligned}$$

Hence the desired conclusion follows from the first isomorphism theorem. □

Theorem 11.62 (Fourth isomorphism theorem)**.** *Let $N \triangleleft G$. The canonical projection homomorphism $G \to G/N$ defines a bijective correspondence between the set of subgroups of G containing N and the set of (all) subgroups of G/N. Under this correspondence normal subgroups correspond to normal subgroups.*

Solvable Groups

Definition 11.63 (Solvable group)**.** A group G is ***solvable*** if there exists a sequence of normal subgroups (known as a *composition series*)

$$G = H_0 \triangleright H_1 \triangleright \cdots \triangleright H_n = \{1\},$$

such that H_i/H_{i+1} (this quotient is called a *composition factor*) is abelian.

Proposition 11.64. *Let $K \triangleleft G$. If K and G/K are solvable, then G is solvable.*

Proof. By definition, and the assumption that K is solvable, it suffices to prove the existence of a sequence of normal subgroups

$$G = H_0 \triangleright H_1 \triangleright \cdots \triangleright H_n = K$$

such that H_i/H_{i+1} is abelian. Let $\bar{G} = G/K$. By assumption, there exists a sequence of normal subgroups

$$\bar{G} = \bar{H}_0 \triangleright \bar{H}_1 \triangleright \cdots \triangleright \bar{H}_n = \{\bar{1}\}$$

such that $\bar{H}_i/\bar{H}_{i+1}$ is abelian.

Consider the quotient map

$$\begin{aligned} \pi \colon G &\to \bar{G} \\ g &\mapsto gK \end{aligned}$$

and let $H_i = \pi^{-1}(\bar{H}_i)$.

Claim. These H_i comprise a composition series for G.

Since the quotient map π is a surjective homomorphism, by 11.61, we have an isomorphism

$$H_i/H_{i+1} \cong \bar{H}_i/\bar{H}_{i+1}$$

and $K = \pi^{-1}(\bar{H}_n)$, so we have found the sequence of subgroups of G as we wanted, proving the result. □

Proposition 11.65. *Let $H \leq G$. If G is solvable, then H is solvable.*

Proof. Since G is solvable, there exists a sequence of normal subgroups

$$G = H_0 \triangleright H_1 \triangleright \cdots \triangleright H_n = \{1\}$$

such that H_i/H_{i+1} is abelian. Consider

$$H = H \cap H_0 \triangleright \cdots \triangleright H \cap H_n = \{1\}.$$

Now $(H \cap H_i) \cap H_{i+1} = H \cap H_{i+1}$. Since $H_i \triangleright H_{i+1}$, this implies $H \cap H_i \triangleright H \cap H_{i+1}$ by looking at the conjugacy relationship. By 11.61,

$$(H \cap H_i)/(H \cap H_{i+1}) \cong (H \cap H_i)H_{i+1}/H_{i+1} \leq H_i/H_{i+1}$$

which is abelian, by the following lemma:

Lemma. *Let G be abelian, and $\phi: G \to G'$ be a surjective homomorphism. Then G' is abelian.*

Proof. Let $x, y \in G'$. We can write $x = \phi(a)$ and $y = g(b)$, by surjectivity of ϕ. Thus $xy = \phi(a)\phi(b) = \phi(ab) = \phi(ba) = \phi(b)\phi(a) = yx$. □

□

Proposition 11.66. *If G is solvable and $\phi: G \to G'$ is a surjective homomorphism, then G' is solvable.*

Proof. Since G is solvable, there exists a sequence of normal subgroups

$$G = H_0 \triangleright H_1 \triangleright \cdots \triangleright H_n = \{1\}$$

such that H_i/H_{i+1} is abelian. Let $H_i' = \phi(H_i)$. Clearly $H_0' = \phi(G) = G'$ (by surjectivity), and $H_n' = \phi(\{1\}) = \{1'\}$.

Let $y \in H_{i+1}'$, and $y = \phi(a)$ for some $a \in H_{i+1}$. Since $H_{i+1} \subset H_i$, $a \in H_i$ and $y \in H_i'$. Thus $H_{i+1}' \leq H_i'$. To show normality, let $x = f(b) \in H_i'$ where $b \in H_i$. Then $xyx^{-1} = \phi(bab^{-1})$. Since $H_{i+1} \triangleleft H_i$, $bab^{-1} \in H_i$, so $\phi(bab^{-1}) \in H_{i+1}'$. Thus $H_{i+1}' \triangleleft H_i'$.

Consider the map

$$\begin{aligned} \psi: H_i/H_{i+1} &\to H_i'/H_{i+1}' \\ hH_{i+1} &\mapsto \phi(h)H_{i+1}' \end{aligned}$$

We claim that ψ is a surjective homomorphism.

1. We first check that ψ is well-defined. If $h_1H_{i+1} = h_2H_{i+1}$, then $h_1h_2^{-1} \in H_{i+1}$, so $\phi(h_1)\phi(h_2)^{-1} \in H_{i+1}'$. Thus $\phi(h_1)H_{i+1}' = \phi(h_2)H_{i+1}'$.

2. ψ is a homomorphism, since

$$\psi(h_1H_{i+1}h_2H_{i+1}) = \psi(h_1h_2H_{i+1}) = \phi(h_1h_2)H_{i+1}' = \phi(h_1)H_{i+1}'\phi(h_2)H_{i+1}' = \psi(h_1H_{i+1})\psi(h_2H_{i+1}).$$

3. If $h'H_{i+1}' \in H_i'/H_{i+1}'$, then since $h' \in H_i'$, $h' = \phi(h)$ for some $h \in H_i$ and $\psi(hH_{i+1}) = h'H_{i+1}'$. Thus ψ is surjective.

By the lemma in the proof of the previous result, H_i'/H_{i+1}' is abelian. □

In a sense, the objects having the "simplest" structure are the building blocks for the more complicated objects. For groups, these are the *simple groups*. All finite simple groups have been classified under the ATLAS of Finite Groups.

Definition 11.67. A group is ***simple*** if it has no non-trivial proper normal subgroup.

That is, G is simple if the only normal subgroups are $\{1\}$ and G.

Example. For prime p, the cyclic group C_p is simple, since it has no proper subgroups at all, let alone normal ones.

Let G be a finite group. Then one can find a sequence of normal subgroups

$$G = H_0 \triangleright H_1 \triangleright \cdots \triangleright H_n = \{1\},$$

such that H_i/H_{i+1} is simple. (This follows from a previous result and the third isomorphism theorem.)

A group may have more than one composition series. However, the Jordan–Hölder theorem states that any two composition series of a given group are equivalent. That is, they have the same composition length and the same composition factors, up to permutation and isomorphism.

Theorem 11.68 (Jordan–Hölder theorem)**.** *Let G be a finite group. Consider two composition series*

$$G = H_0 \triangleright H_1 \triangleright \cdots \triangleright H_n = \{1\}$$

$$G = K_0 \triangleright K_1 \triangleright \cdots \triangleright K_m = \{1\}$$

where each composition factor is simple. Then $n = m$, and the list of composition factors is unique up to permutation.

11.3 Symmetric Groups

Let S be a non-empty set. A bijection $\sigma\colon S \to S$ is called a *permutation* of S; the set of permutations of S is denoted by $\mathrm{Sym}(S)$.

Lemma. $\mathrm{Sym}(S)$ *forms a group under function composition* $\circ$.

We call $\mathrm{Sym}(S)$ the ***symmetric group*** on S.

Proof. If $\sigma\colon S \to S$ and $\tau\colon S \to S$ are bijections, then the composition $\sigma \circ \tau$ is a bijection from S to S. Thus $\circ$ is a binary operation on $\mathrm{Sym}(S)$.

(i) Composition of functions is associative, so $\circ$ is associative.

(ii) The identity of $\mathrm{Sym}(S)$ is the identity map.

(iii) Every bijection has a bijective inverse.

□

In the special case where $S = \{1,2,\ldots,n\} = J_n$, the symmetric group on S is called the *symmetric group of degree* n, and we denote it by S_n.

Lemma 11.69. $|S_n| = n!$

Proof. Obvious, since there are $n!$ permutations of $\{1,2,\ldots,n\}$.

There are n choices for $\sigma(1)$, $n-1$ choices for $\sigma(2)$, ..., 1 choice for $\sigma(n)$. Hence $|S_n| = n(n-1)\cdots 1 = n!$. □

Instead of describing elements of S_n by listing where it sends $1,\ldots,n$, we have more efficient notation.

A *cycle* is a string of integers $(a_1 a_2 \cdots a_m)$ which represents the element of S_n that sends $a_i \mapsto a_{i+1}$ $(i = 1,\ldots,m-1)$, $a_m \mapsto a_1$, and fixes the rest of $1,\ldots,n$.

Example. $(132) \in S_4$ represents the permutation

$$1 \mapsto 3 \quad 3 \mapsto 2 \quad 2 \mapsto 1 \quad 4 \mapsto 4$$

We can write an arbitrary element of S_n as a product of k cycles

$$(a_1 \cdots a_{m_1})(a_{m_1+1} \cdots a_{m_2}) \cdots (a_{m_{k-1}+1} \cdots a_{m_k})$$

called the *cycle decomposition*.

If $\sigma = (a_1 \cdots a_m)$ is a cycle, then one verifies at once that the inverse σ^{-1} is also a cycle, and

$$\sigma^{-1} = (a_m \cdots a_1).$$

To compute a product of cycles, we read from right to left (similar to how one reads the composition of permutations from right to left).

Example. $(132)(34) = (2134)$. One sees this using the definition: If $\sigma = (132)$ and $\tau = (34)$. then

$$\begin{aligned}\sigma(\tau(3)) &= \sigma(4) = 4,\\ \sigma(\tau(4)) &= \sigma(3) = 2,\\ \sigma(\tau(2)) &= \sigma(2) = 1,\\ \sigma(\tau(1)) &= \sigma(1) = 3.\end{aligned}$$

Two cycles are said to be *disjoint* if no number appears in both cycles.

Lemma 11.70. *Disjoint cycles commute.*

Proof. Suppose $\sigma, \tau \in S_n$ are disjoint cycles. We will show that $\sigma(\tau(a)) = \tau(\sigma(a))$.

If a is in neither of σ and τ, then $\sigma(\tau(a)) = \tau(\sigma(a)) = a$.

Otherwise, WLOG assume that a is in τ but not in σ. Then $\tau(a) \in \tau$ and thus $\tau(a) \notin \sigma$. Thus $\sigma(a) = a$ and $\sigma(\tau(a)) = \tau(a)$. Hence we have $\sigma(\tau(a)) = \tau(\sigma(a)) = \tau(a)$.

Therefore τ and σ commute. □

Remove two row notation

There are two ways to denote a permutation (an element of the symmetric group). The first is the *two row notation*: if $\sigma \in S_n$, we write

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & \cdots & n \\ \sigma(1) & \sigma(2) & \sigma(3) & \cdots & \sigma(n) \end{pmatrix}.$$

Lemma 11.71. *S_n is non-abelian for all $n \geq 3$.*

Proof. S_3 consists of

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix},$$

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}.$$

By considering the composition of any two of the above permutations, we see that they do not commute. Thus S_3 is not abelian.

For $n \geq 3$, since we can view S_3 as a subgroup of S_n by fixing $4,5,6,\ldots,n$, it follows that S_n is not abelian. □

Theorem 11.72 (Cayley's theorem). *Every finite group is isomorphic to some subgroup of some symmetric group.*

Proof. Let G be a finite group. For $g \in G$, $\sigma_g(h) = gh$ defines a permutation on G, and $\sigma_{g_1}\sigma_{g_2}(h) = \sigma_{g_1}(g_2h) = g_1g_2h = \sigma_{g_1g_2}(h)$. □

A ***transposition*** τ is a permutation which interchanges two numbers and leaves the others fixed, i.e., there exist distinct $i, j \in J_n$ such that $\tau(i) = j$, $\tau(j) = i$, and $\tau(k) = k$ if $k \neq i$, $k \neq j$.

One sees at once that if τ is a transposition, then $\tau^{-1} = \tau$ and $\tau^2 = \mathrm{id}$. In particular, the inverse of a transposition is a transposition. We shall prove that the transpositions generate S_n.

Proposition 11.73. *Every permutation in S_n can be expressed as a product of transpositions.*

Proof. Induct on n. For $n = 1$, there is nothing to prove since there is only one element. Let $n > 1$ and assume the assertion proved for $n-1$.

Let $\sigma \in S_n$. Let $\sigma(n) = k$. Let $\tau \in S_n$ be such that $\tau(k) = n$, $\tau(n) = k$. Then $\tau\sigma$ is a permutation such that

$$\tau\sigma(n) = \tau(k) = n.$$

In other words, $\tau\sigma$ leaves n fixed. We may therefore view $\tau\sigma$ as a permutation of J_{n-1}, and by induction, there exist transpositions $\tau_1,\ldots,\tau_s \in S_{n-1}$, leaving n fixed, such that

$$\tau\sigma = \tau_1 \cdots \tau_s.$$

We now write

$$\sigma = \tau^{-1}\tau_1 \cdots \tau_s,$$

thereby proving our proposition. □

We shall prove that for $n \geq 5$, the group S_n is not solvable. We need some prehminaries.

Lemma. *Let $H \triangleleft G$. Then G/H is abelian if and only if H contains all elements of the form $xyx^{-1}y^{-1}$, where $x, y \in G$.*

Proof.

$\boxed{\Longrightarrow}$ Consider the quotient map $\pi\colon G \to G/H$. Suppose G/H is abelian. For any $x, y \in G$, we have

$$\pi(xyx^{-1}y^{-1}) = \pi(x)\pi(y)\pi(x)^{-1}\pi(y)^{-1} = H,$$

since G/H is abelian. Hence $xyx^{-1}y^{-1} \in H$.

$\boxed{\Longleftarrow}$ Suppose H contains all elements of the form $xyx^{-1}y^{-1}$, where $x, y \in G$.

Let $\bar{x}, \bar{y} \in G/H$. Since π is surjective, there exist $x, y \in G$ such that $\bar{x} = \pi(x)$, $\bar{y} = \pi(y)$.

Let $\bar{1}$ denote the identity of G/H, and 1 denote the identity of G. Then

$$\bar{1} = \pi(1) = \pi(xyx^{-1}y^{-1}) = \pi(x)\pi(y)\pi(x)^{-1}\pi(y)^{-1} = \bar{x}\bar{y}\bar{x}^{-1}\bar{y}^{-1}.$$

Multiplying by $\bar{y}$ and then $\bar{x}$ on the right, we find

$$\bar{y}\bar{x} = \bar{x}\bar{y}.$$

Hence G/H is abelian. □

Theorem 11.74. *If $n \geq 5$, then S_n is not solvable.*

Proof. We need the following result.

Lemma. *Let $N \triangleleft H \leq S_n$. If H contains every 3-cycle and H/N is abelian, then N contains every 3-cycle.*

Proof. Let i, j, k, r, s be distinct integers in $\{1, \ldots, n\}$, and let

$$\sigma = (ijk), \quad \tau = (krs).$$

Then

$$\begin{aligned} \sigma\tau\sigma^{-1}\tau^{-1} &= (ijk)(krs)(kji)(srk) \\ &= (rki). \end{aligned}$$

Since H contains every 3-cycle, $\sigma, \tau \in H$. Since H/N is abelian, by the above lemma, N contains all elements of the form $\sigma\tau\sigma^{-1}\tau^{-1}$. Thus $\sigma\tau\sigma^{-1}\tau^{-1} \in N$.

Since the choice of i, j, k, r, s was arbitrary, this implies $\sigma\tau\sigma^{-1}\tau^{-1}$ is an arbitrary 3-cycle. Hence N contains every 3-cycle. □

S_n contains all 3-cycles. Thus by induction on the previous lemma, a composition series

$$S_n = H_0 \triangleright H_1 \triangleright H_2 \triangleright \cdots \triangleright H_n$$

must have H_n containing all 3-cycles; thus $H_n \neq \{1\}$ (since the trivial subgroup does not contain any 3-cycles). □

11.4 Group Actions

Group Acting on Sets

We have already seen several examples of groups whose elements are functions: S_n is a set of functions from $\{1,\dots,n\}$ to itself, the elements of D_{2n} are functions from the set of vertices of an n-gon to itself.

The notion of a *group action* generalises this idea.

Definition 11.75 (Group action)**.** Let G be a group, S be a set. An ***action*** of G on S is a map $G \times S \to S$ satisfying the following properties:

(i) $g(hs) = (gh)s$ for all $g,h \in G$, $s \in S$; (associativity)

(ii) $1s = s$ for all $s \in S$. (identity)

Notation. If the group action $\cdot\colon G \times S \to S$ is not clear from context, we write $g \cdot s$ instead of gs.

In fact, each element of G is indeed a function from $S \to S$: for $g \in G$, define $\sigma_g\colon S \to S$ by $s \mapsto gs$.

Lemma 11.76. *If G acts on S and $g \in G$, then σ_g is a permutation of S.*

Proof. We need to show σ_g is bijective.

Consider $\sigma_{g^{-1}}$. If $s \in S$, then

$$\sigma_{g^{-1}}(\sigma_g(s)) = g^{-1}(gs) = 1s = s = g(g^{-1}s) = \sigma_g\left(\sigma_{g^{-1}}(s)\right).$$

Hence $\sigma_{g^{-1}}$ is an inverse for σ_g. □

Note that $\mathrm{Sym}(S)$ is the group consisting of all the permutations of S, so there is a natural relationship between G and $\mathrm{Sym}(S)$:

Lemma 11.77. *If G acts on S, the map $\phi\colon G \to \mathrm{Sym}(S)$ defined by $\phi(g) = \sigma_g$ is a homomorphism.*

Proof. We want to show that for $g_1, g_2 \in G$, $\phi(g_1g_2) = \phi(g_1) \circ \phi(g_2)$. We will show that they are equal on every element of S.

Let $s \in S$. Then

$$\begin{aligned}\phi(g_1g_2)(s) = \sigma_{g_1g_2}(s) = (g_1g_2)(s) = g_1(g_2s) = \sigma_{g_1}(\sigma_{g_2}(s))\\ = \phi(g_1)(\phi(g_2)(s)) = \phi(g_1) \circ \phi(g_2)(s).\end{aligned}$$

□

The *kernel* of the action is

$$\{g \in G \mid gs = s \, \forall s \in S\}.$$

When is this homomorphism injective? Exactly when $\ker \phi = \{1\}$, i.e., 1 is the only element that is the identity on S. (Equivalently, no two distinct group elements induce the same permutation on S.) Such a group action is said to be *faithful*.

Example.

- $D_{2n} \times \{1, \ldots, n\} \to \{1, \ldots, n\}$ is a group action. It is faithful, since only the identity fixed every vertex.
- Consider the vector space $\mathbb{R}^n$. The multiplicative group $\mathbb{R}^\times = \mathbb{R} \setminus \{0\}$ has a natural group action on $\mathbb{R}^n$ given by the vector space structure: if $a \in \mathbb{R}$, then $a(x_1, \ldots, x_n) = (ax_1, \ldots, ax_n)$.
- $GL_n(\mathbf{F})$ acts on $\mathbf{F}^n$ by left multiplication.
- Every group acts on itself by multiplication: $g \cdot s = gs$.

Orbits and Stabilisers

Definition 11.78 (Orbit). Let G act on S. The ***orbit*** of $s \in S$ is

$$Gs := \{gs \mid g \in G\}.$$

Intuitively, it is the elements that s can possibly get mapped to. An element $x \in G_s$ in an orbit is called a *representative* of the orbit, and we say that x *represents* the orbit.

In what follows, the formalism of orbits is similar to the formalism of cosets.

Lemma 11.79. *If $s \in G$, then $Gs = G$.*

Proof. Suppose $s \in G$. Then

$$\begin{aligned} x \in Gs &\iff x = gs, g \in G \\ &\iff x \in G \end{aligned}$$

□

The next result states when two orbits are equal.

Lemma 11.80. *If $t \in Gs$, then $Gt = Gs$.*

Proof. Suppose $t \in G_s$. Then $t = gs$ for some $g \in G$. Then

$$Gt = G(gs) = \{h(gs) \mid h \in G\} = \{(hg)s \mid h \in G\} = (Gg)s = Gs.$$

□

Lemma 11.81. *Suppose the group G acts on S. Then the orbits of the action partition S.*

We need to prove the following:

(i) The orbits are non-empty.

(ii) The union of orbits is S.

(iii) Two orbits are either equal or disjoint.

Proof.

(i) For every $s \in S$, $1s = s$ so $s \in Gs$. Hence every s is in some orbit.

(ii)

(iii) Let $x \in Gs$ and $x \in Gt$. Then $x = g_1 s = g_2 t$ for some $g_1, g_2 \in G$. Thus

$$Gs = Gg_1 s = Gg_2 t = Gt.$$

□

Hence S is a disjoint union of the distinct orbits, and we can write

$$S = \bigcup_{i \in I} Gs_i$$

where I is some indexing set, and the s_i represent distinct orbits.

We say an action G on S is *transitive*, if for all $s \in S$, $Gs = S$. This means that we can reach any element from any element.

Example (Left regular action)**.** Any group G acts on itself by left multiplication:

$$g \cdot s = gs.$$

This action is faithful and transitive.

Proof. If $g, s \in G$, then $g \cdot s = gs \in G$, so the operation is closed. We now show this is an

action.

(i) For all $a \in G$, $1 \cdot a = 1a = a$ by definition of a group.

(ii) For all $g, h \in G$ and $a \in G$, $g(ha) = (gh)a$ by associativity.

To show that it is faithful, we want to show that for all $a \in A$, $g \cdot a = a$ implies $g = 1$; but this follows directly from the uniqueness of identity of the group G.
To show that it is transitive, for all $x, y \in G$, then $(yx^{-1}) \cdot x = y$. Thus any x can be sent to any y. □

Definition 11.82 (Stabiliser)**.** Let G act on S. The ***stabiliser*** of $s \in S$ is

$$G_s := \{g \in G \mid gs = s\}.$$

Intuitively, it is the elements in G that leave s unchanged.

Lemma. $G_s \leq G$.

Proof. Apply the subgroup criterion.

(i) By definition, $1s = s$, so $1 \in G_s$. Thus G_s is non-empty.

(ii) Let $g, h \in G_s$. Then $(gh^{-1})s = g(h^{-1}s) = gs = s$. Thus $gh^{-1} \in G_s$.

□

Since G_s is a subgroup of G, we can consider cosets of G_s in G.

Lemma 11.83. *Let G act on S. If $g, h \in G$ are in the same coset of G_s, then $gs = hs$.*

Proof. If $g, h \in G$ are in the same coset of G_s, then we can write $h = gk$ for some $k \in G_s$. Then

$$hs = (gk)s = g(ks) = gs.$$

□

Theorem 11.84 (Orbit–stabiliser theorem)**.** *Let G act on S, and let $s \in S$. Then there exists a bijection between Gs and cosets of G_s in G. In particular, if G is finite, then*

$$|G| = |Gs|\,|G_s|. \tag{11.6}$$

Proof. We biject the cosets of G_s in G with elements in the Gs. Consider the mapping

$$\begin{aligned} \theta \colon G/G_s &\to Gs \\ gG_s &\mapsto gs \end{aligned}$$

We claim that θ is a bijection.

1. We check that θ is well-defined: Suppose $gG_s = hG_s$. Then $h = gk$ for some $k \in G_s$. Thus $\theta(hG_s) = hs = (gk)s = g(ks) = gs = \theta(gG_s)$.
2. θ is surjective: Let $x \in Gs$. Then there exists $g \in G$ such that $x = gs$. Thus $\theta(gG_s) = gs = x$.
3. θ is injective: Suppose $gs = hs$. Then $h^{-1}gs = s$, so $h^{-1}g \in G_s$. Thus $h^{-1}gG_s = G_s$, which implies $gG_s = hG_s$.

Since θ is a bijection, we have

$$|G/G_s| = |Gs|.$$

Then the result follows from Lagrange's theorem. □

An immediate corollary is a formula for the size of an orbit:

$$|Gs| = |G : G_s|.$$

Suppose S is a finite set. Then we get a decomposition of the order of S as a sum or orders of orbits, which we call the *orbit decomposition formula*:

$$|S| = \sum_{i=1}^{n} |G : G_{s_i}|. \tag{11.7}$$

An important application of the orbit–stabiliser theorem is determining group sizes. To find the order of the symmetry group of, say, a pyramid, we find something for it to act on, pick a favorite element, and find the orbit and stabiliser sizes.

Example. Take $G = D_{2n}$.
Suppose we want to know how big D_{2n} is. D_{2n} acts on the vertices of $\{1,2,\ldots,n\}$ transitively. Since

$$\begin{aligned} |\operatorname{orb}(1)| &= n \\ \operatorname{stab}(1) &= \{e, \text{reflection in the line through } 1\} \end{aligned}$$

we have that $|D_{2n}| = |\operatorname{orb}(1)|\,|\operatorname{stab}(1)| = 2n$.

More Actions

Given any group G, there are a few important actions we can define. In particular, we will define the *conjugation action*, which is a very important concept on its own.

Definition 11.85 (Conjugation of element)**.** The ***conjugation*** of $a \in G$ by $b \in G$ is

$$bab^{-1} \in G.$$

Two elements $a, b \in G$ are *conjugate* if there exists $g \in G$ such that $b = gag^{-1}$.

Lemma 11.86. *Conjugation is an equivalence relation.*

Proof. There are three properties to check:

(i) Since $a = 1a1^{-1}$, a is conjugate to a. (Reflexivity)

(ii) If a is conjugate to b, then $a = gbg^{-1}$ for some $g \in G$, so $b = g^{-1}ag$. (Symmetry)

(iii) Suppose a is conjugate to b, and b is conjugate to c. Then $a = gbg^{-1}$ for some $g \in G$, and $b = hch^{-1}$ for some $h \in G$. Thus $a = (gh)c(gh)^{-1}$. (Transitivity)

□

Lemma 11.87 (Conjugation action)**.** *Any group G acts on itself by conjugation:*

$$g \cdot h = ghg^{-1}$$

for all $g, h \in G$.

Proof. If $g, h \in G$ then $ghg^{-1} \in G$. We now show that this is an action:

(i) $1 \cdot s = 1s1^{-1} = s$.

(ii) $g \cdot (h \cdot k) = g \cdot (hkh^{-1}) = ghkh^{-1}g^{-1} = (gh)k(gh)^{-1} = (gh) \cdot k$.

□

We give special names for the orbits and stabilisers of the conjugation action.

Definition 11.88. The ***conjugacy classes*** are the orbits of the conjugacy action:

$$\mathrm{ccl}(a) := \{b \in G \mid \exists g \in G, gag^{-1} = b\}.$$

The ***centralisers*** are the stabilisers of the conjugation action:

$$C_G(a) := \{g \in G \mid gag^{-1} = a\} = \{g \in G \mid ga = ag\}.$$

By the orbit decomposition formula,

$$|G| = \sum_{i=1}^{n} |G : C_G(g_i)|,$$

where the g_i represent distinct centralisers.

The centraliser is defined as the elements that commute with a particular element h. For the whole group G, we can define the *center*.

Definition 11.89 (Center)**.** The ***center*** of G is the set of elements which commute with all the elements of G:

$$Z(G) := \{g \in G \mid gh = hg \, \forall h \in G\}.$$

Suppose $g_1, \ldots, g_m$ are representatives of the m conjugacy classes which contain more than one element. Note that an element $g \in G$ is in the center of G if and only if the orbit of g is $\{g\}$. In general, the order of the orbit of g is equal to the index of the centraliser of g. Then

$$|G| = |Z(G)| + \sum_{i=1}^{m} |G : C_G(g_i)|. \tag{11.8}$$

This is known as the *class equation*.

In many ways, conjugation is related to normal subgroups.

Lemma 11.90. *Let $H \triangleleft G$. Then G acts by conjugation on H.*

Proof. We only have to prove closure since the other properties follow from the conjugation action. However, by definition of a normal subgroup, for every $g \in G$, $h \in H$, we have $ghg^{-1} \in H$. So it is closed. □

Proposition 11.91. *Normal subgroups are exactly those subgroups which are unions of conjugacy classes.*

Proof. Let $H \triangleleft G$. If $h \in H$, then by definition for every $g \in G$, we get $ghg^{-1} \in H$. So $\mathrm{ccl}(h) \subset H$. So H is the union of the conjugacy classes of all its elements.

Conversely, if H is a union of conjugacy classes and a subgroup of G, then for all $h \in H$, $g \in G$, we have $ghg^{-1} \in H$. So H is normal. □

Lemma 11.92. *Let X be the set of subgroups of G. Then G acts by conjugation on X.*

Proof. We first show closure. If $H \leq G$, we need to show that gHg^{-1} is also a subgroup.

(i) We know that $1 \in H$ and thus $g1g^{-1} = 1 \in gHg^{-1}$, so gHg^{-1} is non-empty.

(ii) For any two elements gag^{-1} and $gbg^{-1} \in gHg^{-1}$, $(gag^{-1})(gbg^{-1})^{-1} = g(ab^{-1})g^{-1} \in gHg^{-1}$.

We now show that it is an action.

(i) $1H1^{-1} = H$.

(ii) $g_1(g_2Hg_2^{-1})g_1^{-1} = (g_1g_2)H(g_1g_2)^{-1}$.

□

Under this action, normal subgroups have singleton orbits.

Definition 11.93 (Normaliser of subgroup)**.** The ***normaliser*** of a subgroup H is the stabiliser of the (group) conjugation action:

$$N_G(H) := \{g \in G \mid gHg^{-1} = H\}.$$

We clearly have $H \subset N_G(H)$. It is easy to show that $N_G(H)$ is the largest subgroup of G in which H is a normal subgroup, hence the name.

There is a connection between actions in general and conjugation of subgroups.

Lemma 11.94. *Stabilisers of the elements in the same orbit are conjugate, i.e., let G act on S and let $g \in G$, $s \in S$. Then $G_{gs} = g\,G_s\,g^{-1}$.*

Applications

Theorem 11.95 (Cauchy's theorem)**.** *Let G be a finite group and prime p dividing $|G|$. Then G has an element of order p (in fact there must be at least $p-1$ elements of order p).*

Proof. Let G and p be fixed. Consider $G^p = G \times \cdots \times G$, the set of p-tuples of G. Let $X \subset G^p$ be

$$X = \{(a_1, \ldots, a_p) \in G^p \mid a_1 \cdots a_p = 1\}.$$

In particular, if an element b has order p, then $(b,b,\ldots,b) \in X$. In fact, if $(b,b,\ldots,b) \in X$ and $b \neq 1$, then b has order p, since p is prime.

Now let $H = \langle h \mid h^p = 1 \rangle \cong C_p$ be a cyclic group of order p with generator h. Let H act on X by "rotation":

$$h(a_1, a_2 \ldots, a_p) = (a_2, a_3, \ldots, a_p, a_1).$$

For closure, if $a_1 \cdots a_p = 1$, then $a_1^{-1} = a_2 \cdots a_p$. So $a_2 \cdots_p a_1 = a^{-1} a_1 = 1$ thus $(a_2, a_3, \ldots, a_p, a_1) \in X$. This is an action:

(i) 1 acts as an identity by construction.

(ii) The associativity condition also works by construction.

As orbits partition X, the sum of all orbit sizes must be $|X|$. We know that $|X| = |G|^{p-1}$ since we can freely choose the first $p-1$ entries and the last one must be the inverse of their product.

Since p divides $|G|$, we see that p also divides $|X|$. We have $|\operatorname{orb}(a1,\ldots,a_p)||\operatorname{stab}_H(a_1,\ldots,a_p)| = |H| = p$. So all orbits have size 1 or p, and they sum to $|X| = p \times$ something. We know that there is one orbit of size 1, namely $(1,1,\ldots,1)$. So there must be at least $p-1$ other orbits of size 1 for the sum to be divisible by p.

In order to have an orbit of size 1, they must look like $(a,a,\ldots,a)$ for some $a \in G$, which has order p. □

Sylow Subgroups

The Sylow theorems are a set of related theorems describing the subgroups of prime power order of a given finite group. They are very powerful, since they can apply to any finite group, and play an important role in the theory of finite groups.

Definition 11.96. Let p be a prime. By a *p-group*, we mean a finite group whose order is a power of p (i.e., p^α for some $\alpha \in \mathbb{N}$).
Subgroups of a group G which are p-groups are called ***p-subgroup*** of G.

Proposition 11.97. *Let G be a non-trivial p-group. Then*

(i) $Z(G)$ is non-trivial;

(ii) G is solvable.

Proof.

(i) By the class equation,

$$|G| = |Z(G)| + \sum |G : G_{x_i}|$$

where the sum is taken over a finite number of elements x_i with $|G : G_{x_i}| \neq 1$.

Since G is a p-group, it follows that p divides $|G|$ and also $|G : G_{x_i}|$. Hence p divides $|Z(G)|$, so the center $Z(G)$ is not trivial.

(ii) $|G/Z(G)|$ divides $|G|$ so $G/Z(G)$ is a p-group, and by (i), we know that $|G/Z(G)| < |G|$. By induction $G/Z(G)$ is solvable. By 11.64, it follows that G is solvable.

□

Definition 11.98. If G is a group of order $p^\alpha m$, where $p \nmid m$, then a subgroup of order p^α is called a ***Sylow p-subgroup*** of G.

Let $Syl_p(G)$ denote the set of Sylow p-subgroups of G, and let $n_p(G)$ denote the number of Sylow p-subgroups of G (or just n_p if G is clear from context).

We shall prove below that such subgroups always exist. For this we need a lemma.

Lemma 11.99. *Let G be a finite abelian group, $|G| = m$. Let p be a prime, $p \mid m$. Then there exists $H \leq G$ such that $|H| = p$.*

Proof. □

The frst Sylow theorem indicates existence of Sylow subgroups, the second Sylow theorem indicates that all Sylow subgroups are related by conjugation, and the third provides constraints on the number of such subgroups.

Theorem 11.100 (Sylow I)**.** *Let G be a finite group, $p \mid |G|$. Then there exists a p-Sylow subgroup of G.*

Proof. □

Theorem 11.101 (Sylow II)**.** *Let G be a finite group.*
If H is a p-subgroup of G, then H is contained in some p-Sylow subgroup.
All p-Sylow subgroups are conjugate.

Theorem 11.102 (Sylow III)**.** *The number of p-Sylow subgroups of G is $\equiv 1 \pmod{p}$.*

Exercises

Exercise 11.1. Let $H, K \leq G$. If $H \cup K \leq G$, show that either $H \subset K$ or $K \subset H$.

Solution. Suppose $H \cup K \leq G$. Suppose, for a contradiction, that $H \not\subset K$ and $K \not\subset H$.

Let $h \in H \setminus K$, $k \in K \setminus H$. Since $H \cup K \leq G$, we have $hk \in H \cup K$.

- Suppose $hk \in H$, and let $h' = hk$. Since $h \in H$ and $H \leq G$, we have $h^{-1} \in H$. Thus $h^{-1}h' = h^{-1}hk = k$. But $h^{-1}h' \in H$ and $k \notin H$, which is a contradiction.
- Suppose $hk \in K$. Then similarly we will arrive at a contradiction.

Therefore, either $H \setminus K = \emptyset$ or $K \setminus H = \emptyset$. Equivalently, $H \subset K$ or $K \subset H$. □

Exercise 11.2. Show that any two cyclic groups of the same order are isomorphic.

Solution. Suppose $\langle x \rangle$ and $\langle y \rangle$ are both cyclic groups of order n. We consider the cases where $n < \infty$ and $n = \infty$.

Case 1: $n < \infty$**.** We claim that the map

$$\begin{aligned} \phi : \langle x \rangle &\to \langle y \rangle \\ x^k &\mapsto y^k \end{aligned}$$

is an isomorphism.

Lemma. *Let G be a group, $g \in G$, let $m, n \in \mathbb{Z}$. Denote $d = \gcd(m,n)$. If $g^n = 1$ and $g^m = 1$, then $g^d = 1$.*

Proof. By Bezout's lemma, since $d = \gcd(m,n)$, then there exists $q, r \in \mathbb{Z}$ such that $qm + rn = d$. Thus

$$g^d = g^{qm+rn} = (g^m)^q (g^n)^r = 1.$$

□

We first show that ϕ is well-defined; that is, $x^r = x^s \implies \phi(x^r) = \phi(x^s)$. Note that $x^{r-s} = e$, so by the above lemma, $n \mid r - s$. Write $r = tn + s$ for some $t \in \mathbb{Z}$, so

$$\phi(x^r) = \phi(x^{tn+s}) = y^{tn+s} = (y^n)^t y^s = y^s = \phi(x^s).$$

We then show that ϕ is a homomorphism:

$$\phi(x^a x^b) = \phi(x^{a+b}) = y^{a+b} = y^a y^b = \phi(x^a)\phi(x^b).$$

Finally we show that ϕ is bijective. Since the element y^k of $\langle y \rangle$ is in the image of x^k under ϕ, ϕ is surjective. Since both groups have the same finite order, any surjection from one to the other is a bijection. Therefore ϕ is an isomorphism.

Case 2: $n = \infty$**.** If $\langle x \rangle$ is an infinite cyclic group, consider the map

$$\begin{aligned} \phi : \mathbb{Z} &\to \langle x \rangle \\ k &\mapsto x^k \end{aligned}$$

(This map is well-defined since there is no ambiguity in the representation of elements in the domain.)

Since $x^a \neq x^b$ for all distinct $a, b \in \mathbb{Z}$, ϕ is injective. By definition of a cyclic group, ϕ is surjective. As above, the laws of exponents ensure ϕ is a homomorphism. Hence ϕ is an isomorphism.

□

Exercise 11.3. The quotient group of a cyclic group is cyclic.

Solution. Let $G = C_n$ with $H \leq C_n$. We know that H is also cyclic; say $C_n = \langle c \rangle$ and $H = \langle c^k \rangle \cong C_\ell$, where $k\ell = n$. We have $C_n/H = \{H, cH, c^2H, \ldots, c^{k-1}H\} = \langle cH \rangle \cong C_k$. □

Exercise 11.4. Every subgroup of index 2 is normal.

Solution. Suppose $H \leq G$ has index 2.

Claim. $G/H = \{H, G \setminus H\}$.

G/H is the set of left cosets of H in G. Since H has index 2 in G, there must be 2 cosets of H in G; one of these cosets must be H itself.

Since cosets form a partition of the group, the other coset must be all the remaining elements of G that are not in H, which is $G \setminus H$. Hence $G/H = \{H, G \setminus H\}$.

By the same reasoning, the only two right cosets are H and $G \setminus H$.

Hence all left cosets and right cosets are the same, so H is normal. □

Exercise 11.5. A group of order 6 is either cyclic or dihedral.

Solution. Let $|G| = 6$. We will show that either $G \cong C_6$ or $G \cong D_6$.

By Lagrange's theorem, the possible element orders are 1, 2, 3 and 6. If there exists $a \in G$ of order 6, then $G = \langle a \rangle \cong C_6$.

Otherwise, we can only have elements of orders 2 and 3 other than the identity. If G only has elements of order 2, the order must be a power of 2 (why), which is not the case. So there must be an element r of order 3. So $\langle r\rangle \triangleleft G$ as it has index 2. Now G must also have an element s of order 2 (why).

Since $\langle r\rangle$ is normal, we know that $srs^{-1} \in \langle r\rangle$. If $srs^{-1} = 1$, then $r = 1$, which is not true. If $srs^{-1} = r$, then $sr = rs$ and sr has order 6 (lcm of the orders of s and r), which was ruled out above. Otherwise if $srs^{-1} = r^2 = r^{-1}$, then G is dihedral by definition of the dihedral group. □

Exercise 11.6. Any non-abelian group of order 6 is isomorphic to S_3.

From this classification theorem we obtain $D_6 \cong S_3$ and $GL_2(\mathbf{F}_2) \cong S_3$, without havng to find explicit maps between the groups.

Solution. By Cauchy's theorem, a group of order 6 has an element x of order 2, and an element y of order 3. These two elements generate the group. The 6 elements e, y, y^2, x, xy, xy^2 must all be different from each other, hence this is the list of all elements of the group. Therefore, yx must be somewhere on this list.

Checking each element: we know that $yx \neq e$ because $x \neq y^{-1}$, $yx \neq y$ because $x \neq e$, $yx \neq y^2$ because $x \neq y$, $yx \neq x$ because $y \neq e$, and $yx \neq xy$ because by assumption the group is not abelian. Thus $yx = xy^2$, hence our group is the symmetric group S_3. □

Exercise 11.7 (NUS MA2202S AY23/24)**.** Let G be a group. Suppose G has a unique subgroup H of order n. Prove that H is a normal subgroup of G.

Solution. Let $g \in G$. Consider the map

$$\begin{aligned} \phi : G &\to gHg^{-1} \\ h &\mapsto ghg^{-1} \end{aligned}$$

which is bijective, so $|gHg^{-1}| = |H| = n$.

But $|H|$ is the unique subgroup of order n, so $gHg^{-1} = H$. Hence H is a normal subgroup of G. □

Exercise 11.8 (NUS MA2202S AY23/24)**.** Let G be a p-group for some prime p with a normal subgroup H. Assume H is of order p. Prove that H is contained in the center of G.

Solution. Let $|G| = p^n$ for $n \in \mathbb{N}$.

Note that $N_G(H)/C_G(H) = G/C_G(H)$ is isomorphic to a subgroup of $\mathrm{Aut}(H)$ which has order $p-1$. Since $p^n/|C_G(H)|$ divides $p-1$, $p^n/|C_G(H)|$ must be a power of p less than or equal to

$p-1$, implying

$$p^n/|C_G(H)| = 1 \implies |C_G(H)| = p^n = |G| \implies C_G(H) = G.$$

Hence, all elements of H commute with all elements of G, thus, $H \subset Z(G)$. □

12 Rings

12.1 Rings

Definitions and Properties

Definition 12.1 (Ring). A ***ring*** R is a set together with two binary operations $+$ and $\times$ (called addition and multiplication), satisfying the following axioms:

(i) $(R,+)$ is an abelian group, with identity 0;

(ii) $\times$ is associative: $(a\times b)\times c = a\times(b\times c)$ for all $a,b,c\in R$;

(iii) $\times$ distributes over $+$: for all $a,b,c\in R$,

$$a\times(b+c) = (a\times b)+(a\times c),$$
$$(a+b)\times c = (a\times c)+(b\times c).$$

Notation. We simply write ab rather than $a\times b$, for $a,b\in R$.

Notation. Denote the additive identity of $a\in R$ by $-a$.

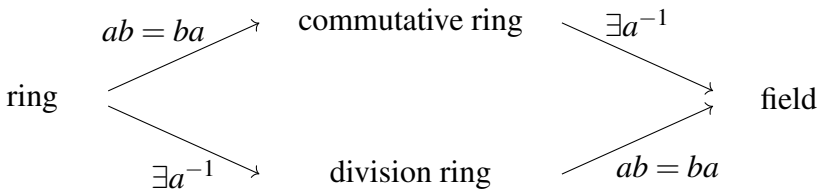
We say R is *commutative* if multiplication is commutative.

We say R has an *identity* if there exists $1\in R$ such that

$$1\times a = a\times 1 = a \quad (a\in R).$$

In general, a ring may not necessarily be commutative or have multiplicative inverses; when they do, we give such rings special names.

Definition 12.2. A ring R with identity 1, where $1\neq 0$, is called a ***division ring*** if every $a\in R\setminus\{0\}$ has a multiplicative inverse, i.e., exists $b\in R$ such that $ab = ba = 1$.
A commutative division ring is called a ***field***.



Example.

- $\mathbb{Z}$ is the prototypical ring; it is not a field.
- $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ are fields.
- $\mathbb{Z}/n\mathbb{Z}$ is a commutative ring with identity $\bar{1}$ under addition and multiplication of residue classes.
- $2\mathbb{Z}$ is a commutative ring without identity.
- The trivial ring $R = \{0\}$ is a commutative ring with identity $1 = 0$.
- The ring of (polynomial/continuous/differentiable) functions on $\mathbb{R}$.
- The endomorphism ring $\mathrm{End}_{\mathbb{R}}(V)$ of a vector space V over $\mathbb{R}$ is a non-commutative ring.
- The Hamilton Quaternions $\mathbb{H}$. Historically the first example of a non-commutative ring.

Lemma 12.3 (Basic properties). *Let R be a ring.*

(i) $0a = a0 = 0$ *for all* $a \in R$.

(ii) $(-a)b = a(-b) = -(ab)$ *for all* $a, b \in R$.

(iii) $(-a)(-b) = ab$ *for all* $a, b \in R$.

(iv) If R has identity 1*, then the identity is unique and* $-a = (-1)a$.

Proof. These all follow from the distributive laws and cancellation in the additive group $(R, +)$.

(i) We have $0a = (0+0)a = 0a + 0a$. Then the cancellation law implies $0a = 0$.
Similarly, $a0 = a(0+0) = a0 + a0$. Thus $a0 = 0$.

(ii) We have $ab + (-a)b = (a + (-a))\, b = 0b = 0$. Thus $(-a)b = -(ab)$.
Similarly, $ab + a(-b) = a\,(b + (-b)) = a0 = 0$. Thus $a(-b) = -(ab)$.

(iii) Using (ii), $(-a)(-b) = -a(-b) = -\,(-(ab)) = ab$.

(iv) Suppose 1 and $1'$ are identities of R. Then $1 = 1 \times 1' = 1'$.

Since $(-1)a + a = (-1)a + 1a = ((-1)+1)a = 0a = 0$, it follows that $-a = (-1)a$.

□

Subrings

Having defined the notion of a ring, there is a natural notion of a subring.

Definition 12.4 (Subring). Let R be a ring. We say $S \subset R$ is a ***subring*** of R, if S is a subgroup of R that is closed under multiplication.

Example.

- $\mathbb{Z}$ is a subring of $\mathbb{Q}$, and $\mathbb{Q}$ is a subring of $\mathbb{R}$.
- $n\mathbb{Z} = \{nk \in \mathbb{Z} \mid k \in \mathbb{Z}\}$ is a subring of $\mathbb{Z}$.
- The real-valued differentiable functions on $\mathbb{R}$ form a subring of the ring of continuous functions.
- $\mathbb{Z}[i] = \{x + yi \mid x, y \in \mathbb{Z}\}$ is a subring of $\mathbb{C}$, called the ring of Guassian integers.
- $\mathbb{Q}[\sqrt{2}] = \{x + y\sqrt{2} \mid x, y \in \mathbb{Q}\}$ is a subring of $\mathbb{R}$.
- $S = \mathbb{Z} + \mathbb{Z}i + \mathbb{Z}j + \mathbb{Z}k$ form a subring of $\mathbb{H}$.

We will use the square brackets notation quite frequently. It should be clear what it should mean, and we will define it properly later.

If R contains 1, then S is a (unital) subring if $1_R \in S$. We assume subrings are unital unless otherwise specified.

Lemma 12.5 (Subring criterion). *Let R be a ring, $S \subset R$. Then S is a subring of R if and only if*

(i) $1 \in S$; *(non-empty)*

(ii) $ab, a - b \in S$ *for all* $a, b \in S$. *(closed under multiplication and subtraction)*

Proof.

$\boxed{\Longrightarrow}$

$\boxed{\Longleftarrow}$ The condition that $a - b \in S$ for all $a, b \in S$ implies that S is an additive subgroup by the subgroup test (note that as $1 \in S$ we know that S is nonempty). The other conditions for a

subring hold directly. □

Units, Zero Divisors

Recall that in a ring we do not require that non-zero elements have a multiplicative inverse. Nevertheless, since the multiplication operation is associative and there is a multiplicative identity, the elements which happen to have multiplicative inverses form a group:

Definition 12.6. Let R be a ring, with identity $1 \neq 0$. We say $u \in R$ is a ***unit*** in R if there exists $v \in R$ such that $uv = vu = 1$.
We say $a \in R \setminus \{0\}$ is a ***zero divisor*** if there exists $b \in R \setminus \{0\}$ such that either $ab = 0$ or $ba = 0$.

Remark. A zero divisor can not be a unit.

Let $R^\times$ denote the set of units in R.

Lemma. *$R^\times$ forms a group under multiplication.*

We call $R^\times$ the ***group of units***.

Proof.

(i) Evidently $1 \in R^\times$.

(ii) Let $u_1, u_2 \in R^\times$. Then $u_1 v_1 = 1$, $u_2 v_2 = 1$ for some $v_1, v_2 \in R$. Thus $(u_1 u_2)(v_2 v_1) = 1$. Similarly $(v_2 v_1)(u_1 u_2) = 1$. Hence $u_1 u_2 \in R^\times$.

(iii) Let $u \in R^\times$. Then $uv = vu = 1$ for some $v \in R$. Taking inverse gives $u^{-1}v^{-1} = v^{-1}u^{-1} = 1$. Hence $u^{-1} \in R^\times$.

□

Example.

- The ring $\mathbb{Z}$ has no zero divisors and its only units are ± 1.
- The group of units of $\mathbb{Z}/n\mathbb{Z}$ is $(\mathbb{Z}/n\mathbb{Z})^\times$. Recall that $(\mathbb{Z}/n\mathbb{Z})^\times = \{a \in \mathbb{Z}/n\mathbb{Z} \mid (a,n) = 1\}$. All elements not in $(\mathbb{Z}/n\mathbb{Z})^\times$ are zero divisors. In sum, every non-zero element of $\mathbb{Z}/n\mathbb{Z}$ is either a unit or a zero divisor.

Rings having some of the same characteristics as $\mathbb{Z}$ are given a name:

Definition 12.7 (Integral domain)**.** If a commutative ring with identity $1 \neq 0$ has no zero divisors, it is called an ***integral domain***.

Example.

- $\mathbb{Z}$ is an integral domain.
- All fields are integral domains.

The absence of zero divisors in integral domains give these rings a cancellation property:

Lemma 12.8.

(i) Let $a,b,c \in R$, a is not a zero divisor. If $ab = ac$, then either $a = 0$ or $b = c$.

(ii) In particular, for any a,b,c in an integral domain and $ab = ac$, then either $a = 0$ or $b = c$.

Proof.

(i) If $ab = ac$, then $a(b-c) = 0$. Since a is not a zero divisor, we have either $a = 0$ or $b - c = 0$.

(ii) This follows from (i) and the definition of an integral domain.

□

Corollary 12.9. *Any finite integral domain is a field.*

In this terminology, a field is a commutative ring F with identity $1 \neq 0$ in which every non-zero element is a unit, i.e., $F^\times = F \setminus \{0\}$.

Proof. Let R be a finite integral domain, let $a \in R \setminus \{0\}$.

By the cancellation law, the map $x \mapsto ax$ is an injective function. Since R is finite, this map is also surjective. In particular, there exists $b \in R$ such that $ab = 1$, i.e., a is a unit in R. Since a was an arbitrary non-zero element, R is a field. □

Corollary 12.10. *If p is a prime, $\mathbb{Z}/p\mathbb{Z}$ is a field, usually denoted by $\mathbf{F}_p$.*

Examples

Example (Matrix rings)**.** Let R be a (often commutative) ring with 1. We define the matrix ring $M_{n\times n}(R)$ as the set consisting of

$$(a_{ij})_{n\times n}, \quad a_{ij} \in R.$$

Addition and multiplication on $M_{n\times n}(R)$ is defined following the matrix multiplication in linear algebra.
If we take $R=\mathbb{R}$, then $M_{n\times n}(\mathbb{R})$ the usual matrix algebra. We have the subring of diagonal matrices, and the subring of upper triangular matrices.

Example (Group rings)**.** Let R be a commutative ring with 1. Let G be a finite group. We define the group ring $R[G]$ as the set consisting of

$$\sum_{g\in G} a_g g \quad (a_g \in R).$$

Addition on $R[G]$ is defined in the obvious/naive way. The multiplication is via the following example

$$(a_g g + a_h h)(a_{g'} g' + a_{h'} h') = a_g a_{g'} gg' + a_h a_{g'} hg' + a_g a_{h'} gh' + a_h a_{h'} hh',$$

where gg', hg', gh', hh' are the group multiplication in G.

Lemma. *Let R be a commutative ring with 1. Let G be a finite group.*

(i) Let $e \in G$ be the identity element. Then 1_e is the identity of the ring $R[G]$.

(ii) Let $e \neq g \in G$. Then $1-g$ is a zero divisor.

(iii) Let H be a subgroup of G. Then $R[H]$ is a subring of $R[G]$.

(iv) The ring $R[G]$ is commutative if and only if G is commutative.

Example (Product of rings)**.** Let R and S be two rings. We define the ring $R\times S$ as follows: as a set $R\times S$ is the same as the Cartesian product of sets; we define the addition and multiplication component wise:

$$(a,b)+(c,d) = (a+c, b+d),$$
$$(a,b)\times(c,d) = (ac, bd).$$

12.2 Homomorphisms and Isomorphisms

Ideals

Definition 12.11 (Ideal). Let R be a ring. We say $I \subset R$ is a ***left ideal*** if

(i) $(I,+)$ is a subgroup of $(R,+)$; (additive subgroup)

(ii) $ax \in I$ for all $a \in R$, $x \in I$. (closed under left multiplication)

We define a ***right ideal*** similarly.
We say I is a (two-sided) ***ideal*** of R, if I is both a left ideal and a right ideal of R.

We say I is a *proper ideal* if $I \neq R$.

Remark. For commutative rings, left ideals, right ideals, and (two-sided) ideals coincide

Example.

- Trivial ideals: the zero ideal $\{0\}$ and the whole ring R are two-sided ideals.

  R is also called the *unit ideal*: if $x \in R^\times \cap I$, then $x^{-1}x = 1 \in I$, so $a \times 1 = a \in I$ for all $a \in R$. Thus $I = R$. (We have shown $I = R$ if and only if $1 \in I$.)

  This implies that in a field F, the only ideals are $\{0\}$ and F, since if $I \neq \{0\}$, let $x \in F \setminus \{0\}$, then x is a unit, so $I = F$.

- The even numbers $2\mathbb{Z} = (2)$ is an ideal of $\mathbb{Z}$.

The next definition provides a way to generate an ideal from an element of a ring.

Definition 12.12 (Principal ideal). Let R be a ring, and let $a \in R$. The ***principal left ideal*** generated by a is

$$(a) := \{xa \mid x \in R\}.$$

More generally, let $a_1, \ldots, a_n \in R$. Define

$$(a_1, \ldots, a_n) := \{x_1a_1 + \cdots + x_na_n \mid x_i \in R\}.$$

We call $a_1, \ldots, a_n$ *generators* for this ideal.

Lemma. *$(a_1, \ldots, a_n)$ is a left ideal.*

Proof. If $y_1,\dots,y_n,x_1,\dots,x_n \in R$ then

$$\begin{aligned}(x_1a_1+\cdots+x_na_n)+(y_1a_1+\cdots+y_na_n) &= x_1a_1+y_1a_1+\cdots+x_na_n+y_na_n\\ &= (x_1+y_1)a_1+\cdots+(x_n+y_n)a_n.\end{aligned}$$

If $z \in R$, then

$$z(x_1a_1+\cdots+x_na_n) = zx_1a_1+zx_na_n.$$

Finally,

$$0 = 0a_1+\cdots+0a_n.$$

□

Example. Let R be a ring. Let L,M be left ideals. Define the product

$$LM = \{x_1y_1+\cdots+x_ny_n \mid x_i \in L, y_i \in M\}.$$

Then LM is also a left ideal.
If L,M,N are left ideals, then $(LM)N = L(MN)$.

Example. Let L,M be left ideals. Define the sum

$$L+M = \{x+y \mid x \in L, y \in M\}.$$

Then $L+M$ is a left ideal.
If L,M,N are left ideals, then $L(M+N) = LM+LN$.

Example. The ideals of $\mathbb{Z}$ are (n) for $n \in \mathbb{N}$, and $\{0\}$.

Proof. Let $I \neq \{0\}$ be an ideal of $\mathbb{Z}$. Let $a \in I$ be non-zero. Since $a,-a \in I$, I contains a natural number. By well-ordering, there is a minimal $n \in \mathbb{N} \cap I$.
Clearly $(n) \subset I$, since all multiples of n are contained in I. If $x \in I$ and $x \neq (n)$, by the division algorithm, we can write $x = qn+r$ for some $q,r \in \mathbb{Z}$, $0 < r < n$. But $qn \in I$, so $x-qn = r \in I$. Then r is a smaller natural number in I, which contradicts the minimality of n. Thus $I = (n)$. □

This shows that $\mathbb{Z}$ is a principal ideal domain (all ideals of $\mathbb{Z}$ are principal ideals).

Homomorphisms

Definition 12.13. We say $\phi : R \to S$ is a ***homomorphism*** if it satisfies

(i) $\phi(a+b) = \phi(a) + \phi(b)$ for all $a, b \in R$;

(ii) $\phi(ab) = \phi(a)\phi(b)$ for all $a, b \in R$;

(iii) $\phi(1_R) = 1_S$.

An ***isomorphism*** is a bijective homomorphism. Two rings R and S are ***isomorphic***, denoted by $R \cong S$, if there exists an isomorphism between R and S.

Remark. For groups, condition (iii) is not required in the definition: $\phi(1)\phi(x) = \phi(1x) = \phi(x)$ then we can cancel on both sides due to the existence of (multiplicative) inverse.

An isomorphism between a ring with itself is called an *automorphism*.

An injective homomorphism $\phi : R \to S$ is called an ***embedding***[1]; we say R is *embedded* in S.

Definition 12.14. Let $\phi : R \to S$ be a homomorphism. The ***kernel*** of ϕ is its kernel viewed as a homomorphism of additive groups:

$$\ker \phi := \{r \in R \mid \phi(r) = 0\}.$$

The ***image*** of ϕ is

$$\operatorname{im} \phi := \{s \in S \mid \exists r \in R, \phi(r) = s\}.$$

Example.

- Consider the quotient map $\pi : \mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$; $\ker \pi = n\mathbb{Z}$.
- The embedding of the subring $n\mathbb{Z} \to \mathbb{Z}$. The kernel is trivial.
- The map

$$\begin{aligned} \phi : \mathbb{C}[x] &\to \mathbb{C} \\ f(x) &\mapsto f(a) \end{aligned}$$

  The kernel is

$$\ker \phi = \{f(x) \in \mathbb{C}[x] \mid f(a) = 0\} = \{(x-a)f(x) \mid f(x) \in \mathbb{C}[x]\}.$$

[1] If ϕ is injective, then $R \cong \operatorname{im} \phi$, where $\operatorname{im} \phi$ is a subring of S, so we can think of R as a "subring" of S; hence the term *embedding* to mean that R is "contained in" S.

Lemma 12.15. *Let $\phi : R \to S$ be a homomorphism. Then*

(i) $\ker\phi$ *is a ideal of R;*

(ii) $\operatorname{im}\phi$ *is a subsring of S.*

Proof.

(i) Let $x, y \in \ker\phi$. Then

$$\phi(x-y) = \phi(x) - \phi(y) = 0 - 0 = 0$$

so $x - y \in \ker\phi$. Thus $\ker\phi$ is an additive subgroup of R.

Let $r, r' \in R$, $x \in \ker\phi$. Then

$$\phi(rxr') = \phi(r)\phi(x)\phi(r') = \phi(r)0\phi(r') = 0.$$

Thus $rxr' \in \ker\phi$, and so $\ker\phi$ is an ideal of R.

(ii)

□

Lemma 12.16. *Let $\phi : R \to S$ be a homomorphism. Then ϕ is injective if and only if* $\ker\phi = \{0\}$.

Proof. This follows from considering $(R, +)$ as an additive group. Then the result follows from group theory. □

Quotient Rings

Let $I \subset R$ be an ideal, and let $a \in R$. Define

$$a + I = \{a + x \mid x \in I\}.$$

This is usually not an ideal, but rather an *additive coset* of I (considering I as an additive subgroup of R). Any element of the coset is called a *representative* of the coset.

Definition 12.17 (Quotient ring)**.** Let $I \subset R$ be an ideal. Then the ***quotient ring*** is

$$R/I := \{a + I \mid a \in R\}.$$

Lemma. *R/I is a ring, with addition and multiplication defined as*

$$(a+I)+(b+I) = (a+b)+I,$$
$$(a+I)\cdot(b+I) = ab+I.$$

Proof. Recall that by 11.50, a subgroup of an abelian group is normal. Since I is an additive subgroup of R, and $(R,+)$ is abelian, we have $I \triangleleft R$ under addition. Hence the quotient group $(R/I,+)$ is defined.

We now check that multiplication is well-defined. Suppose $a+I = a'+I$, $b+I = b'+I$. Then $a-a' = r \in I$, $b-b' = s \in I$. Thus

$$ab = (a'+r)(b'+s) = a'b' + a's + b'r + rs.$$

Note that $a's, b'r, rs \in I$. Hence $ab+I = a'b'+I$.

Check that R/I is a ring, with additive identity 0_R+I and multiplicative identity 1_R+I. □

Example. Take $R = \mathbb{Z}$, $I = (n)$ for some $n \in \mathbb{N}$. We can write $(n) = n\mathbb{Z}$, so the quotient ring is $\mathbb{Z}/n\mathbb{Z}$.

As before, we give a name to the canonical homomorphism from R to R/I.

Definition 12.18 (Quotient map). Let $I \subset R$ be an ideal. The ***quotient map*** is

$$\pi\colon R \to R/I$$
$$a \mapsto a+I$$

Lemma 12.19. *Quotient maps are surjective homomorphisms.*

Proof. Let $\pi\colon R \to R/I$ be a quotient map.

- Let $a,b \in R$. Then $\pi(a+b) = (a+b)+I = (a+I)+(b+I) = \pi(a)+\pi(b)$.
- Let $a,b \in R$. Then $\pi(ab) = ab+I = (a+I)(b+I) = \pi(a)\pi(b)$.
- $\pi(1_R) = 1_R+I$, which is the identity of R/I.

□

In addition,

$$\ker\pi = \{a \in R \mid a+I = 0_R+I\} = \{a \in R \mid a \in I\} = I.$$

Isomorphism Theorems

Theorem 12.20 (First isomorphism theorem). *Let $\phi : R \to S$ be a homomorphism. Then*

$$R/\ker\phi \cong \operatorname{im}\phi. \tag{12.1}$$

Proof. Denote $K = \ker\phi$. Consider the map

$$\begin{aligned} \theta : R/K &\to \operatorname{im}\phi \\ a+K &\mapsto \phi(a) \end{aligned}$$

We claim that θ is an isomorphism.

1. We first check that θ is well-defined. If $a+K = a'+K$, then $a-a' \in K$, so $\phi(a-a') = 0$. Thus $\phi(a) = \phi(a')$.
2. θ is a homomorphism:

$$\begin{aligned} \theta((a+K)+(b+K)) &= \theta((a+b)+K) = \phi(a+b) = \phi(a)+\phi(b) = \theta(a+K)+\theta(b+K) \\ \theta((a+K)(b+K)) &= \theta(ab+K) = \phi(ab) = \phi(a)\phi(b) = \theta(a+K)\theta(b+K) \\ \theta(0_R+I) &= \phi(0_R) = 0_S \end{aligned}$$

3. θ is injective: $\theta(a+K) = \theta(b+K) \implies \phi(a) = \phi(b) \implies a+K = b+K$.
4. θ is surjective: Let $x \in \operatorname{im}\phi$. Then $x = \phi(a)$ for some $a \in R$. Thus $\theta(a+K) = \phi(a) = x$.

□

Theorem 12.21 (Second isomorphism theorem). *Let A be a subring, and B be an ideal of R. Then*

$$(A+B)/B \cong A/(A\cap B). \tag{12.2}$$

Lemma.

(i) $A+B = \{a+b \mid a \in A, b \in B\}$ *is a subring of* R;

(ii) $A \cap B$ *is an ideal of* A.

Theorem 12.22 (Third isomorphism theorem). *Let I and J be ideals of R, with $I \subset J$. Then*

$$(R/I)(J/I) \cong R/J. \tag{12.3}$$

Lemma. *J/I is an ideal of R/I.*

Theorem 12.23 (Fourth isomorphism theorem)**.** *Let I be an ideal of R. The correspondance $A \leftrightarrow A/I$ is an inclusion preserving bijection between the set of subrings of A of R that contain I and the set of subrings of R/I. Furthermore, A (a subring containing I) is an ideal of R if and only if A/I is an ideal of R/I.*

Chinese Remainder Theorem

Definition 12.24. Let R be a commutative ring. We say two ideals $I, J \subset R$ are *coprime* if

$$I + J = R.$$

In particular, there $i \in I$, $j \in J$ such that $i + j = 1$.

Theorem 12.25 (Chinese remainder theorem)**.** *Let R be a commutative ring. Suppose I and J are coprime ideals of R. Then for any $a, b \in R$, there exists $x \in R$ such that*

$$x \in (a + I) \cap (b + J).$$

Proof. Let $i \in I$ and $j \in J$ be such that $i + j = 1$.

Claim. $x = aj + bi$.

We can write

$$x = a(1 - i) + bi = a + (b - a)i \in a + I.$$

Similarly,

$$x = aj + b(1 - j) = b + (a - b)j \in b + J.$$

□

modular arithmetic

Prime and Maximal Ideals

Let R be a commutative ring.

Definition 12.26. An ideal $P \subsetneq R$ is ***prime*** if $ab \in P$ implies either $a \in P$ or $b \in P$.
An ideal $M \subsetneq R$ is ***maximal*** if there is no ideal between M and R, i.e., $M \subset I \subset R$ implies $I = M$ or $I = R$.

Example. In $\mathbb{Z}$, (p) is a prime ideal for prime p.
Further $p\mathbb{Z} \subset U = n\mathbb{Z} \subset \mathbb{Z}$, and $p \in U$, then $p = nq$ for some $q \in \mathbb{Z}$. But p is prime and $n \neq 1$ so $n = p$. Thus $U = p\mathbb{Z}$. Thus $p\mathbb{Z}$, for p prime, is a maximal ideal in $\mathbb{Z}$. Note that $0 \subset p\mathbb{Z} \subset \mathbb{Z}$, so 0 is not a maximal ideal in $\mathbb{Z}$.

Lemma 12.27. *Let R be a commutative ring.*

(i) A maximal ideal is prime.

(ii) An ideal P is prime if and only if R/P is integral.

(iii) An ideal M is maximal if and only if R/M is a field.

Proof.

(i) Suppose M is a maximal ideal. Let $ab \in M$, WLOG assume $a \notin M$. Then $M \subsetneq (a) + M = R$, since M is a maximal ideal.

Thus $xa + m = 1$ for some $x \in R$, $m \in M$. Then $b = xab + mb \in M$, since $ab, m \in M$. Hence M is prime.

□

Characteristic of Ring

In the following, let $R = \{0\}$ be a ring; let e denote the identity of R (to distinguish it from the identity of $\mathbb{Z}$). For any $a \in R$, $n \in \mathbb{Z}$, we can define an integer multiple of a ring element:

$$na = \begin{cases} \underbrace{a + \cdots + a}_{n \text{ times}} & (n > 0) \\ -(ka) & (n < 0, n = -k) \\ 0 & (n = 0) \end{cases}$$

Consider the map

$$\begin{aligned} f \colon \mathbb{Z} &\to R \\ n &\mapsto ne \end{aligned}$$

Then this is a homomorphism (this is a bit tedious, since we have to consider $n > 0$, $n < 0$ or $n = 0$). Now let $f \colon \mathbb{Z} \to R$ be any homomorphism. By definition, $f(1) = e$. Then if $n > 0$, $f(n) = f(1 + \cdots + 1) = f(1) + \cdots + f(1) = nf(1) = ne$. Hence there is one and only one homomorphism $\mathbb{Z} \to R$.

Assume $R \neq \{0\}$. Let $f\colon \mathbb{Z} \to R$ be *the* homomorphism. Since $\ker f$ is an ideal of $\mathbb{Z}$, $\ker f = n\mathbb{Z}$ for some integer $n \geq 0$. (Note that $n \neq 1$, otherwise $\ker f = \mathbb{Z}$ so $\operatorname{im} f = \{0\}$, but $f(1) = e \neq 0$.)

By the first isomorphism theorem, $\mathbb{Z}/n\mathbb{Z} \cong \operatorname{im} f$. In practice, we do not make any distinction between $\mathbb{Z}/n\mathbb{Z}$ and its image in R, and we agree to say that "R contains $\mathbb{Z}/n\mathbb{Z}$ as a subring".

Suppose $n \neq 0$. Then for all $a \in R$,

$$\underbrace{a + \cdots + a}_{n \text{ times}} = na = (ne)a = f(n)a = 0a = a.$$

We call n the ***characteristic*** of R, or say R has characteristic n, and denote $n = \operatorname{char}(R)$.

Remark. If $n = 0$, then $\mathbb{Z}/0\mathbb{Z} = \mathbb{Z}$, so rings of characteristic 0 are infinite (since it contains a subring isomorphic to $\mathbb{Z}$, which is infinite).

Note that $n \neq 0$ is the smallest positive integer m such that $me = 0$. This is because $m \in \ker f$, so $n \mid m$, which implies $n \leq m$.

Lemma 12.28. *Suppose R is an integral ring. Then $\operatorname{char}(R)$ is either 0 or prime.*

Proof. Suppose $n = \operatorname{char}(R) \neq 0$. Suppose, for a contradiction, that n is composite. Then $n = mk$, where $m, k > 1$. Then $m, k < n$.

By minimality of n, we have $me, ke \neq 0$. But $(me)(ke) = mke = ne = 0$. This implies that R has zero divisors, which contradicts the assumption that R is an integral ring. □

Lemma 12.29 (Freshman's dream)**.** *Let R be commutative with prime characteristic p. Then $(x+y)^p = x^p + y^p$ for all $x, y \in R$.*

Proof. Since R is commutative, we have the binomial expansion:

$$(x+y)^p = \sum_{i=1}^{p} \binom{p}{i} x^i y^{p-i}.$$

(We require R to be commutative, so that we can freely move variables around in order to raise them by powers.) For $i \in \{1, \ldots, p-1\}$, $\binom{p}{i}$ is divisible by p. Since $\operatorname{char}(R) = p$, multiples of p equal 0. Hence $(x+y)^p = x^p + 0 + \cdots + 0 + y^p = x^p + y^p$. □

Let K be a field, and let $f\colon \mathbb{Z} \to K$ be the homomorphism from the integers to K. If $\ker f = \{0\}$, then K contains $\mathbb{Z}$ as a subring, and we say that K has *characteristic* 0. If $\ker f = p\mathbb{Z}$ for some prime p, then we say K has *characteristic* p.

The field $\mathbb{Z}/p\mathbb{Z}$ is sometimes denoted by $\mathbf{F}_p$, and is called the *prime field*, of characteristic p. This prime field $\mathbf{F}_p$ is contained in every field of characteristic p.

Quotient Fields

Recall that we can construct $\mathbb{Q}$ from $\mathbb{Z}$, using equivalence classes of ordered pairs whose elements are in $\mathbb{Z}$. Instead of $\mathbb{Z}$, our discussion will apply to an arbitrary integral ring R.

Let $(a,b),(c,d) \in R \times R^*$, where $R^* = R \setminus \{0\}$; we call these ordered pairs *quotients*. Define a relation $R \times R^*$:

$$(a,b) \sim (c,d) \iff ad = bc.$$

Lemma. *$\sim$ is an equivalence relation on $R \times R^*$.*

Proof.

(i) Since $ab = ba$, we have $(a,b) \sim (a,b)$.

(ii) Suppose $(a,b) \sim (c,d)$. Then $ad = bc$, or $cb = da$. This implies $(c,d) \sim (a,b)$.

(iii) Suppose $(a,b) \sim (c,d)$ and $(c,d) \sim (e,f)$. Then

$$ad = bc, \quad cf = de.$$

Thus

$$adf = bcf = bde,$$

so $daf - dbe = 0$. Then $d(af - be) = 0$. Since R has no divisors of 0, and $d \neq 0$, it follows that $af - de = 0$, i.e., $af = be$. This means that $(a,b) \sim (e,f)$.

□

We denote the equivalence class of (a,b) by a/b; that is,

$$\frac{a}{b} = \{(c,d) \in R \times R^* \mid (a,b) \sim (c,d)\}.$$

Then the ***quotient field*** (or *field of fractions*) of R is the set of equivalence classes:

$$\mathrm{Frac}(R) := (R \times R^*)/\sim$$

with addition and multiplication defined by

$$\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd},$$
$$\frac{a}{b}\frac{c}{d} = \frac{ac}{bd}.$$

Lemma 12.30. $\mathrm{Frac}(R)$ *is a field, with addition and multiplication being defined as above.*

Proof. We first check that addition and multiplication, as defined above, are well-defined.

Addition Suppose $a/b = a'/b'$ and $c/d = c'/d'$. We must show that

$$\frac{ad+bc}{bd} = \frac{a'd'+b'c'}{b'd'}.$$

This is true if and only if

$$b'd'(ad+bc) = bd(a'd'+b'c'),$$

or in other words,

$$b'd'ad + b'd'bc = bda'd' + bdb'c'.$$

But $ab' = a'b$ and $cd' = c'd$ by assumption. Hence the above equation holds.

Multiplication Suppose $a/b = a'/b'$ and $c/d = c'/d'$.

The verification that $\mathrm{Frac}(R)$ is a commutative ring with identity is left as an exercise; note that the additive identity is $0/1$, and the multiplicative identity is $1/1$ (where 1 is the identity of R).

We now show that $\mathrm{Frac}(R)$ is a field. Note that if $a/b = 0/1$, then $(a,b) \sim (0,1)$, so $a = 0$. Thus if $a/b \neq 0/1$ is a non-zero element, then $a \neq 0$. Then (b,a) and subsequently b/a is well-defined (since $a \neq 0$). The multiplicative of a/b is then b/a:

$$\frac{a}{b}\frac{b}{a} = \frac{b}{a}\frac{a}{b} = \frac{ab}{ab} = \frac{1}{1}.$$

Hence every non-zero element in $\mathrm{Frac}(R)$ has a multiplicative inverse, so $\mathrm{Frac}(R)$ is a field. □

Example.

- $\mathbb{Q} = \mathrm{Frac}(\mathbb{Z})$.
- $\mathbb{Q}[i] = \mathrm{Frac}(\mathbb{Z}[i])$, the field of Gaussian rationals.
- The quotient field of a field is canonically isomorphic to the field itself.

Lemma 12.31. *R is embedded in* $\mathrm{Frac}(R)$.

Proof. Consider the map

$$\begin{aligned} \phi : R &\to \mathrm{Frac}(R) \\ a &\mapsto a/1 \end{aligned}$$

We claim that ϕ is an embedding (injective homomorphism).

1. ϕ is a homomorphism:

$$\begin{aligned}\phi(a+b) &= \frac{a+b}{1} = \frac{a}{1} + \frac{b}{1} = \phi(a) + \phi(b)\\ \phi(ab) &= \frac{ab}{1} = \frac{a}{1}\frac{b}{1} = \phi(a)\phi(b)\\ \phi(1) &= \frac{1}{1}\end{aligned}$$

2. ϕ is injective: $\phi(a) = \phi(b) \implies a/1 = b/1 \implies a = b$.

□

We often think of rationals as an integer dividing another non-zero integer, instead of considering them as equivalence classes. We now show this more generally.

Lemma 12.32. *Suppose R is a subring of a field F. (Thus R is an integral domain.) Then*

$$\mathrm{Frac}(R) \cong \{ab^{-1} \mid a,b \in R, b \neq 0\}.$$

Proof. We see that $\{ab^{-1} \mid a,b \in R, b \neq 0\}$ is a field, which is a subfield of F. Consider the map

$$a/b \mapsto ab^{-1}.$$

We claim this is an isomorphism. □

Hence we often call the field $\{ab^{-1} \mid a,b \in R, b \neq 0\}$ the *quotient field* of R in F; there can be no confusion with this terminology due to the above isomorphism. In view of this, the element ab^{-1} of F is also denoted by a/b.

Proposition 12.33. *Let $\phi : R \to F$ be an embedding of an integral domain R into a field F. Then there exists a unique extension $\phi^* : \mathrm{Frac}(R) \to F$ which is also an embedding. (ϕ^* being an extension of ϕ means $\phi^*|_R = \phi$.)*

Proof.

Existence Define the map

$$\begin{aligned}\phi^* : \mathrm{Frac}(R) &\to F\\ \frac{a}{b} &\mapsto \frac{\phi(a)}{\phi(b)}\end{aligned}$$

1. We first check that ϕ^* is well-defined. Suppose $a/b = c/d$. Then $ad = bc$, so $\phi(ad) = \phi(bc)$, or $\phi(a)\phi(d) = \phi(b)\phi(c)$, which implies $\frac{\phi(a)}{\phi(b)} = \frac{\phi(c)}{\phi(d)}$. (Note that $b,d \neq 0$, so $\phi(b), \phi(d) \neq 0$, since $\ker\phi = \{0\}$ due to injectivity.)

2. ϕ^* is a homomorphism:

$$\begin{aligned}\phi^*\left(\frac{a}{b}+\frac{c}{d}\right) &= \phi^*\left(\frac{ad+bc}{bd}\right) = \frac{\phi(ad+bc)}{\phi(bd)} = \frac{\phi(a)\phi(d)+\phi(b)\phi(c)}{\phi(b)\phi(d)} \\ &= \frac{\phi(a)}{\phi(b)}+\frac{\phi(c)}{\phi(d)} = \phi^*\left(\frac{a}{b}\right)+\phi^*\left(\frac{c}{d}\right) \\ \phi^*\left(\frac{a}{b}\frac{c}{d}\right) &= \phi^*\left(\frac{ac}{bd}\right) = \frac{\phi(ac)}{\phi(bd)} = \frac{\phi(a)\phi(c)}{\phi(b)\phi(d)} = \frac{\phi(a)}{\phi(b)}\frac{\phi(c)}{\phi(d)} = \phi^*\left(\frac{a}{b}\right)\phi^*\left(\frac{c}{d}\right) \\ \phi^*\left(\frac{1}{1}\right) &= \frac{\phi(1)}{\phi(1)} = \frac{1}{1} = 1\end{aligned}$$

3. ϕ^* is injective: Let $a/b \in \ker\phi^*$. Then $\phi^*(a/b) = \phi(a)/\phi(b) = 0$, so $\phi(a) = 0$. By injectivity, $a = 0$, since $\ker\phi = \{0\}$. This implies $a/b = 0/1$, so $\ker\phi^* = \{0\}$.

4. ϕ^* is an extension of ϕ: Since $\phi(1) = 1$, we have $\phi^*(a/1) = \phi(a)/1 = \phi(a)$ for all $a \in R$.

Uniqueness Suppose we have yet to define ϕ^* as above. Then

$$\phi^*\left(\frac{a}{b}\right) = \phi^*\left(\frac{a}{1}\frac{1}{b}\right) = \phi^*\left(\frac{a}{1}\right)\phi^*\left(\frac{1}{b}\right) = \phi^*\left(\frac{a}{1}\right)\phi^*\left(\frac{b}{1}\right)^{-1} = \phi(a)\phi(b)^{-1}.$$

Hence there is only one map ϕ^*, defined as above, which satisfies the above conditions. □

12.3 Euclidean Domains, Principal Ideal Domains, and Unique Factorisation Domains

In this section, we focus our attention on integral domains. We shall cover the following three types of integral domains:

$$\text{Euclidean domains} \subset \text{Principal ideal domains} \subset \text{Unique factorisation domains}.$$

Euclidean Domains

Definition 12.34 (Euclidean domain)**.** An integral domain R is called a ***Euclidean domain*** if there exists $d : R \setminus \{0\} \to \mathbb{Z}_{\geq 0}$ that satisfies: for all $a, b \in R$, $b \neq 0$, there exists $q, r \in R$ such that $a = bq + r$ and $r = 0$ or $d(r) < d(b)$.

Example.

- Consider $\mathbb{Z}$. Let $a, b \in \mathbb{Z}$, $b \neq 0$. Then $a = bq + r$, $d : \mathbb{Z} \setminus \{0\} \to \mathbb{Z}_{\geq 0}$ and $d(x) = |x|$.
- Let F be a field, $F[x]$ be the ring of polynomials with elements of F as coefficients. Consider long division. $d : F[x] \setminus \{0\} \to \mathbb{Z}_{\geq 0}$ and $d(f(x)) := \deg f$.
- Consider $\mathbb{Z}[i] = \{a + bi \mid a, b \in \mathbb{Z}\}$ where $i^2 = -1$. Then $\mathbb{Z}[i]$ is an integral domain with unit $1 = 1 + 0i$ Then $d : \mathbb{Z}[i] \setminus \{0\} \to \mathbb{Z}_{\geq 0}$ under $d(a + bi) = a^2 + b^2$.

Theorem 12.35. *Let R be a Euclidean domain, I be an ideal in R. Then there exists $a_0 \in R$ such that $I = Ra_0$, i.e., I is a principal ideal.*

Principal Ideal Domains

Definition 12.36 (Principal ideal domain)**.** A ***principle ideal domain*** (PID) is an integral domain in which every ideal is a principle ideal.

Example. A field is a PID. The ring $\mathbb{Z}$ is a PID (actually a Euclidean domain).

Proposition 12.37. *Every Euclidean domain is a PID.*

Proposition 12.38. *Every field is a Euclidean domain.*

Unique Factorisation Domains

Gauss's lemma and Eisenstein's criterion for irreducibility.

12.4 Polynomial Rings

Polynomials and Polynomial Functions

Let R be a commutative ring. Define the ***polynomial ring***

$$R[t] := \{a_0 + a_1 t + \cdots + a_n t^n \mid a_i \in R\}.$$

That is, $R[t]$ is the set of polynomials in t with coefficients in R.

Remark. A rigorous definition of the polynomial ring can be found in [Lan05].

Let R be a subring of a commutative ring S. If $f \in R[t]$ is a polynomial, then we may define the associated *polynomial function*

$$f_S \colon S \to S$$

by letting for $x \in S$

$$f_S(x) = f(x) = a_0 + a_1 x + \cdots + a_n x^n.$$

Hence f_S is a function (mapping) from S to itself, determined by the polynomial f.

Given $x \in S$, there is a homomorphism $R[t] \to S$ which maps $f \mapsto f_S(x)$. (show why)

Greatest Common Divisor

Unique Factorisation

Partial Fractions

Polynomials Over Rings and Over the Integers

Principal Rings and Factorial Rings

Polynomials in Several Variables

Symmetric Polynomials

The Mason–Stothers Theorem

The *abc* Conjecture

13 Modules and Vector Spaces

13.1 Modules

Basic Definitions and Examples

We start with the definition of a *module*.

Definition 13.1 (Module). Let R be a ring. A ***R-left module*** (or a left module over R) is an abelian group M together with an action of R on M such that for any $r,s \in R$ and $m,n \in M$,

(i) $(r+s)m = rm+sm$;

(ii) $(rs)m = r(sm)$;

(iii) $r(m+n) = rm+rs$;

(iv) $1m = m$, if R has 1.

Right modules are defined similarly. Unless explicitly mentioned otherwise, the term "module" will always mean "left module".

Example.

- We have the 0 module for any ring.
- Let F be a field. Then F-modules are just F-vector spaces.
- Let M be an abelian group. Then M is a $\mathbb{Z}$-module.
- Let R be a ring. Then R is a left R-module via left multiplication, and a right R-module via right multiplication.
- Let I be a left ideal of R. Then I is a left R-module. Actually I is a left R-submodule of R.

- Let I be a left (or right) ideal of R. Then R/I (the quotient abelian group) is a left (or right) R-module.

Definition 13.2 (Submodule). Let M be a R-module. Then a subgroup $N \subset M$ is called ***R-submodule*** of M, if N is closed under the R-action.

Submodules of M are therefore just subsets of M which are themselves modules under the restricted operations.

Example. Every R-module M has thw two submodules M and 0 (the latter is called the *trivial submodule*).

We establish a submodule criterion analogous to that for subgroups of a group.

Lemma 13.3 (Submodule criterion). *Let R be a ring, and let M be an R-module. Then $N \subset M$ is a submodule of M if and only if*

(i) $N \neq \emptyset$;

(ii) $x + ry \in N$ *for all* $r \in R$, $x, y \in N$.

We then collect some basic results on modules.

Lemma 13.4 (Basic properties). *Let R be with* 1*, and let M be a R-module.*

Homomorphisms and Isomorphisms

Definition 13.5. Let R be a ring, and let M and N be R-modules. We say $\phi : M \to N$ is an ***R-module homomorphism*** if

(i) $\phi(x+y) = \phi(x) + \phi(y)$ for $x, y \in M$; (additivity)

(ii) $\phi(rx) = r\phi(x)$ for $r \in R$, $x \in M$. (homogeneity)

Note that $\phi : M \to N$ is an R-module homomorphism if and only if $\phi(rx+y) = r\phi(x) + \phi(y)$ for all $x, y \in M$, $r \in R$.

Proof.

$\boxed{\Longrightarrow}$ Suppose $\phi : M \to N$ is an R-module homomorphism. Then certainly $\phi(rx+y) = r\phi(x) + \phi(y)$.

$\boxed{\Longleftarrow}$ Suppose $\phi(rx+y) = r\phi(x)+\phi(y)$. Take $r = 1$ to see that ϕ is additive, and take $y = 0$ to see that ϕ commutes with the action of R on M (i.e., is homogeneous). □

Definition 13.6. A bijective R-module morphism $\phi : M \to N$ is called an ***isomorphism***. The modules M and N are ***isomorphic***, denoted $M \cong N$, if there exists an R-module isomorphism $\phi : M \to N$.

We denote $\mathrm{Hom}_R(M,N)$ as the set of R-module morphisms from M to N. We often write $\mathrm{End}_R(M) = \mathrm{Hom}_R(M,M)$, called the *endomorphism ring* of M; its elements are called *endomorphisms*.

Definition 13.7. Let $\phi : M \to N$ be a R-module homomorphism. Then we define the ***kernel*** and ***image*** of ϕ respectively as

$$\ker\phi := \{m \in M \mid \phi(m) = 0\},$$
$$\mathrm{im}\,\phi := \phi(M) = \{n \in N \mid \exists m \in M,\ \phi(m) = n\}.$$

Lemma 13.8. *Let $\phi : M \to N$ be a module homomorphism. Then*

(i) $\ker\phi$ *is a submodule of M;*

(ii) $\mathrm{im}\,\phi$ *is a submodule of N.*

Proof.

(i)

(ii)

□

Lemma 13.9. *Let M, N and L be R-modules.*

(i) Let $\phi, \psi \in \mathrm{Hom}_R(M,N)$. Then $\phi + \psi \in \mathrm{Hom}_R(M,N)$ and with this operation $\mathrm{Hom}_R(M,N)$ is an abelian group.

For $r \in R$, $r\phi \in \mathrm{Hom}_R(M,N)$ and with this action of the commutative ring R the abelian group $\mathrm{Hom}_R(M,N)$ is an R-module.

(ii) If $\phi \in \mathrm{Hom}_R(L,M)$ and $\psi \in \mathrm{Hom}_R(M,N)$, then $\phi \circ \psi \in \mathrm{Hom}_R(L,N)$.

(iii) With addition as above and multiplication defined as function composition, $\mathrm{Hom}_R(M,N)$ is a ring with 1. When R is commutative, $\mathrm{Hom}_R(M,N)$ is an R-algebra.

Proof.

(i)

(ii)

(iii)

□

Theorem 13.10 (First isomorphism theorem)**.** *Let $\phi : M \to N$ be an R-module homomorphism. Then*

$$M/\ker\phi \cong \operatorname{im}\phi.$$

Theorem 13.11 (Second isomorphism theorem)**.** *Let A, B be submodules of the R-module M. Then*

$$(A+B)/B \cong A/(A\cap B).$$

Theorem 13.12 (Third isomorphism theorem)**.** *Let M be an R-module, and let A and B be submodules of M with $A \subset B$. Then*

$$(M/A)/(B/A) \cong M/B.$$

Theorem 13.13 (Fourth isomorphism theorem)**.** *Let N be a submodule of the R-module M.*

Generation of Modules, Direct Sums, and Free Modules

Tensor Products of Modules

13.2 Vector Spaces

13.3 Modules Over Principal Ideal Domains

IV
Real Analysis

Real analysis deals with the real numbers and real-valued functions of a real variable.

- Chapter 14: This chapter defines the real numbers $\mathbb{R}$, the complex numbers $\mathbb{C}$, and Euclidean space $\mathbb{R}^k$.
- Chapter 15: Defining a notion of distance on a set gives rise to a metric space. This chapter discusses various structures in a metric space. We then
- Chapter 16: We study the behaviour of sequences and series.
- Chapter 17:
- Chapter 18:
- Chapter 19:
- Chapter 20:

14 Real and Complex Number Systems

14.1 Ordered Sets and Boundedness

Definitions

Let S be a set.

Definition 14.1 (Order). An ***order*** on S is a binary relation $<$ such that

(i) for all $x, y \in S$, exactly one of $x < y$, $x = y$, or $y < x$ holds; (trichotomy)

(ii) if $x, y, z \in S$ are such that $x < y$ and $y < z$, then $x < z$. (transitivity)

We write $x \leq y$ if $x < y$ or $x = y$. We define $>$ and $\geq$ in the obvious way.

Definition 14.2 (Ordered set). An ***ordered set*** is a set in which an order is defined.

Example. $\mathbb{Q}$ is an ordered set if $r < s$ is defined to mean that $s - r$ is a positive rational number.

Definition 14.3 (Boundedness). Let $E \subset S$, where S is an ordered set.

(i) We say β is an *upper bound* of E if $x \leq \beta$ for all $x \in E$; if E has an upper bound, we say E is ***bounded above***.

(ii) We say β is a *lower bound* of E if $x \geq \beta$ for all $x \in E$; if E has a lower bound, we say E is ***bounded below***.

If E is bounded above and below, we say E is ***bounded***.

A set may have multiple upper and lower bounds. We give special names to the *least* upper bound and the *greatest* lower bound.

Definition 14.4. We say $\alpha \in S$ is the ***supremum*** (or *least upper bound*) of E if

(i) α is an upper bound for E;

(ii) if $\beta < \alpha$, then β is not an upper bound of E (i.e., $\exists x \in S$ such that $x > \beta$).

Similarly, we say $\alpha \in S$ is the ***infimum*** (or *greatest lower bound*) of E if

(i) α is a lower bound for E;

(ii) if $\beta > \alpha$ then β is not a lower bound of E (i.e., $\exists x \in S$ such that $x < \beta$).

Lemma (Uniqueness of suprenum)**.** *If E has a supremum, then it is unique.*

Proof. Suppose α are β be suprema of E.

Since β is a supremum, it is an upper bound for E. Since α is a supremum, then it is the *least* upper bound, so $\alpha \leq \beta$. Interchanging the roles of α and β gives $\beta \leq \alpha$. Hence $\alpha = \beta$. □

We denote the supremum of E by $\sup E$, the infimum by $\inf E$.

Example. Let $E = \left\{\frac{1}{n} \,\middle|\, n \in \mathbb{N}\right\}$. Then $\sup E = 1$, $\inf E = 0$.

Proof. It is clear that 1 is an upper bound for E. Suppose $\beta < 1$. Since $1 \in E$, evidently β is not an upper bound for E. Hence $\sup E = 1$.
It is clear that 0 is a lower bound for E. Suppose $\beta > 0$. Pick $n = \left\lfloor \frac{1}{\beta} \right\rfloor + 1$, then $\beta > \frac{1}{n}$, so β is not a lower bound for E. Hence $\inf E = 0$. □

This shows that the supremum and infimum of a set may not belong to the set itself.

Least-upper-bound Property

Definition 14.5. An ordered set S has the ***least-upper-bound property*** (l.u.b.) if every non-empty subset of S that is bounded above has a supremum in S.

We define the *greatest-lower-bound property* similarly.

Proposition 14.6. *Suppose S is an ordered set. If S has the least-upper-bound property, then S has the greatest-lower-bound property.*

Proof. Let non-empty $B \subset S$ be bounded below. We want to show that $\inf B \in S$.

Let $L \subset S$ be the set of all lower bounds of B:

$$L := \{y \in S \mid y \leq x \quad \forall x \in B\}.$$

Since B is bounded below, B has a lower bound, so $L \neq \emptyset$. Since every $x \in B$ is an upper bound of L, L is bounded above. By the l.u.b. property of S, $\sup L \in S$.

Claim. $\inf B = \sup L$.

To show that $\sup L = \inf B$ (greatest lower bound), we need to show that (i) $\sup L$ is a lower bound of B, (ii) and $\sup L$ is the greatest of the lower bounds.

(i) Suppose $\gamma < \sup L$, then γ is not an upper bound of L. Since B is the set of upper bounds of L, $\gamma \notin B$. Considering the contrapositive, if $\gamma \in B$, then $\gamma \geq \sup L$. Hence $\sup L$ is a lower bound of B, and thus $\sup L \in L$.

(ii) If $\beta > \sup L$ then $\beta \notin L$, since $\sup L$ is an upper bound of L. In other words, $\sup L$ is a lower bound of B, but β is not if $\beta > \sup L$. This means that $\sup L$ is the greatest of the lower bounds.

Hence $\inf B = \sup L \in S$. □

Corollary 14.7. *If S has the greatest-lower-bound property, then it has the least-upper-bound property.*

Hence S has the least-upper-bound property if and only if S has the greatest-lower-bound property.

Properties of Suprema and Infima

There is a corresponding set of properties of the infimum that the reader should formulate for himself.

The next result shows that a set with a supremum contains numbers arbitrarily close to its supremum.

Lemma 14.8 (Approximation property). *Let S be non-empty, $b = \sup S$. Then for every $a < b$ there exists $x \in S$ such that*

$$a < x \leq b.$$

Proof. We first show $x \leq b$. Since $b = \sup S$ is an upper bound of S, $x \leq b$ for all $x \in S$.

We now show there exist $x \in S$ such that $a < x$. Suppose otherwise, for a contradiction, that $x \leq a$ for every $x \in S$. Then a would be an upper bound for S. But since $a < b$ and b is the supremum, this means a is smaller than the least upper bound, a contradiction. □

For the rest of this section, suppose S has the least-upper-bound property.

Lemma 14.9 (Additive property). *Given non-empty $A, B \subset S$, let*

$$C = \{x + y \mid x \in A, y \in B\}.$$

If each of A and B has a supremum, then C has a supremum, and

$$\sup C = \sup A + \sup B.$$

Proof. Let $a = \sup A$, $b = \sup B$. Let $z \in C$, then $z = x + y$ for some $x \in A$, $y \in B$. Then

$$z = x + y \leq a + b,$$

so $a + b$ is an upper bound for C. Since C is non-empty and bounded above, by the lub property of S, C has a supremum in S.

Let $c = \sup C$. To show $a + b = c$, we will show (i) $a + b \geq c$, and (ii) $a + b \leq c$.

(i) Since c is the *least* upper bound for C, and $a + b$ is an upper bound for C, we must have that $c \leq a + b$.

(ii) Choose any $\varepsilon > 0$. By 14.8, there exist $x \in A$, $y \in B$ such that

$$a - \varepsilon < x, \quad b - \varepsilon < y.$$

Adding these inequalities gives

$$a + b - 2\varepsilon < x + y \leq c.$$

Thus $a + b < c + 2\varepsilon$ for every $\varepsilon > 0$. Hence $a + b \leq c$.

□

Lemma 14.10 (Comparison property). *Let non-empty $A, B \subset S$ such that $a \leq b$ for every $a \in A$, $b \in B$. If B has a supremum, then A has a supremum, and*

$$\sup A \leq \sup B.$$

Proof. Let $\beta = \sup B$. Then $b \leq \beta$ for all $b \in B$.

For each $a \in A$, choose any $b \in B$. Then $a \leq b \leq \beta$. Thus β is an upper bound for A.

Since A is non-empty and bounded above, by the lub property of S, A has a supremum in S; let $\alpha = \sup A$. Since β is an upper bound for A, and α is the *least* upper bound for A, we have that $\alpha \leq \beta$, as desired. □

Lemma 14.11. *Let $B \subset S$ be non-empty and bounded below. Let*

$$A = -B := \{-b \mid b \in B\}.$$

Then A is non-empty and bounded above. Furthermore, $\inf B$ *exists, and* $\inf B = -\sup A$.

Proof. Since B is non-empty, so is A. Since B is bounded below, let β be a lower bound for B. Then $b \geq \beta$ for all $b \in B$, which implies $-b \leq -\beta$ for all $b \in B$. Hence $a \leq -\beta$ for all $a \in A$, so $-\beta$ is an upper bound for A.

Since A is non-empty and bounded above, by the lub property of S, A has a supremum. Then $a \leq \sup A$ for all $a \in A$, so $b \geq -\sup A$ for all $b \in B$. Thus $-\sup A$ is a lower bound for B.

Also, we saw before that if β is a lower bound for B then $-\beta$ is an upper bound for A. Then $-\beta \geq \sup A$ (since $\sup A$ is the least upper bound), so $\beta \leq -\sup A$. Therefore $-\sup A$ is the greatest lower bound of B. □

Fields

Recall that a *field* is a commutative division ring:

Definition 14.12 (Field)**.** A field is a set F with operations of addition and multiplication which satisfy the following properties:

(A1) Addition is commutative: $x+y = y+x$ for all $x, y \in F$.

(A2) Addition is associative: $(x+y)+z = x+(y+z)$ for all $x, y, z \in F$.

(A3) Additive identity: there exists $0 \in F$ such that $0+x = x$ for every $x \in F$.

(A4) Additive inverse: tor every $x \in F$, there exists $-x \in F$ such that $x+(-x) = 0$.

(M1) Multiplication is commutative: $xy = yx$ for all $x, y \in F$.

(M2) Multiplication is associative: $(xy)z = x(yz)$ for all $x, y, z \in F$.

(M3) Multiplicative identity: there exists $1 \in F$, $1 \neq 0$ such that $1x = x$ for every $x \in F$.

(M4) Multiplicative inverse: for each $x \in F$, $x \neq 0$, there exists $1/x \in F$ such that $x(1/x) = 1$.

(D) Distributive law: $x(y+z) = xy+xz$ for all $x, y, z \in F$.

It is easy to verify that the additive and multiplicative identity of a field are unique, and each element has a unique additive and multiplicative identity.

Lemma 14.13. *For every $x \in F$, $x0 = 0$.*

Proof. Let $x \in F$. Then $x0 = x(0+0) = x0 + x0$. Now add $-(x0)$ to both sides to get $0 = x0$. □

The familiar properties of arithmetic all follow easily from the field properties listed in the definition of a field. For example, here are a few properties of the additive inverse.

Lemma 14.14.

(i) $-(-x) = x$ *for each* $x \in F$.

(ii) $(-1)x = -x$ *for each* $x \in F$.

(iii) analogous properties for the multiplicative inverse in a field.

Definition 14.15 (Ordered field)**.** A field F is an ***ordered field*** if there exists an order $<$ on F such that for all $x, y, z \in F$,

(i) if $y < z$ then $x + y < x + z$;

(ii) if $x > 0$ and $y > 0$ then $xy > 0$.

If $x > 0$, we call x *positive*; if $x < 0$, x is *negative*.

All the familiar rules for working with inequalities apply in every ordered field: Multiplication by positive [negative] quantities preserves [reverses] inequalities, no square is negative, etc. The following result lists some of these.

Lemma 14.16 (Basic properties)**.** *Let F be an ordered field, $x, y, z \in F$.*

(i) If $x > 0$ then $-x < 0$, and vice versa.

(ii) If $x > 0$ and $y < z$ then $xy < xz$.

(iii) If $x < 0$ and $y < z$ then $xy > xz$.

(iv) If $x \neq 0$ then $x^2 > 0$. In particular, $1 > 0$.

(v) If $0 < x < y$ then $0 < \frac{1}{y} < \frac{1}{x}$.

Proof.

(i) If $x > 0$ then $0 = -x + x > -x + 0$, so that $-x < 0$.

If $x < 0$ then $0 = -x + x < -x + 0$, so that $-x > 0$.

(ii) Since $z > y$, we have $z - y > y - y = 0$, so $x(z-y) > 0$. Hence

$$xz = x(z-y) + xy > 0 + xy = xy.$$

(iii) By (i) and (ii),

$$-[x(z-y)] = (-x)(z-y) > 0,$$

so that $x(z-y) < 0$. Hence $xz < xy$.

(iv) If $x > 0$, part (ii) of the above definition gives $x^2 > 0$.

If $x < 0$, then $-x > 0$ so $(-x)^2 > 0$. But $x^2 = (-x)^2$.

Since $1 = 1^2$, $1 > 0$.

(v) If $y > 0$ and $v \leq 0$, then $yv \leq 0$. But $y\left(\frac{1}{y}\right) = 1 > 0$, so $\frac{1}{y} > 0$. Likewise, $\frac{1}{x} > 0$.

Multiplying both sides of the inequality $x < y$ by the positive quantity $\left(\frac{1}{x}\right)\left(\frac{1}{y}\right)$, we obtain $\frac{1}{y} < \frac{1}{x}$.

□

14.2 Real Numbers

Problems with $\mathbb{Q}$

$\mathbb{Q}$ has some problems, the first of which being *algebraic incompleteness*: there exists equations with coefficients in $\mathbb{Q}$ but do not have solutions in $\mathbb{Q}$ (in fact $\mathbb{R}$ has this problem too, but $\mathbb{C}$ is algebraically complete, by the fundamental theorem of algebra).

Lemma 14.17. *$x^2 - 2 = 0$ has no solution in $\mathbb{Q}$.*

Proof. Suppose, for a contradiction, that $x^2 - 2 = 0$ has a solution $x = \frac{p}{q}$, $q \neq 0$. We also assume $\frac{p}{q}$ is in lowest terms, i.e., p and q are coprime.

Squaring both sides gives $\frac{p^2}{q^2} = 2$, or $p^2 = 2q^2$. Observe that p^2 is even, so p is even; let $p = 2m$ for some integer m. This then implies $4m^2 = 2q^2$, or $2m^2 = q^2$. Similarly, q^2 is even so q is even.

Since p and q share a common factor of 2, we have reached a contradiction. □

Proposition 14.18. *Define*

$$A := \{p \in \mathbb{Q} \mid p > 0,\ p^2 < 2\},$$
$$B := \{p \in \mathbb{Q} \mid p > 0,\ p^2 > 2\}.$$

A contains no largest number, and B contains no smallest number.

Proof. For each rational $p > 0$, let

$$q := p - \frac{p^2 - 2}{p + 2} = \frac{2p + 2}{p + 2}.$$

Thus

$$q^2 - 2 = \frac{2(p^2 - 2)}{(p + 2)^2}.$$

If $p \in A$, $p^2 < 2$ implies $q > p$ and $q^2 - 2 < 0 \implies q \in A$. Thus A has no largest number.

If $p \in B$, $p^2 > 2$ implies $q < p$ and $q^2 - 2 > 0 \implies q \in B$. Thus B has no smallest number. □

Remark. The formula for p might seem to be "rabbit-out-of-hat". We motivate it as such: If $p^2 < 2$ we want to increase p slightly, while if $p^2 > 2$ we want to decrease it, so the amount we should change it by should be obtained from $p^2 - 2$. A denominator is needed to prevent overshooting, especially when p is large, so we use one that grows with p; the actual choice of denominator $p + 2$ can be regarded as the result of trial and error.

Corollary 14.19. *$\mathbb{Q}$ does not have the least-upper-bound property.*

Proof. In the previous result, note that B is the set of all upper bounds of A, and B does not have a smallest element.

Hence $A \subset \mathbb{Q}$ is bounded above but A has no least upper bound in $\mathbb{Q}$. □

Similarly, B is bounded below: The set of all lower bounds of B consists of A and of all $r \in \mathbb{Q}$ with $r \leq 0$. Since A has no largest member, B has no greatest lower bound in $\mathbb{Q}$.

The second problem that $\mathbb{Q}$ has is *analytic incompleteness*: there exists a sequence in $\mathbb{Q}$ which converges to a point that is not in $\mathbb{Q}$; for example, the sequence (a_n) defined by $a_n = \frac{1}{10^{n-1}} \left\lfloor 10^{n-1}\sqrt{2} \right\rfloor$

$$1, 1.4, 1.41, 1.414, 1.4142, \ldots$$

converges to $\sqrt{2}$, which is irrational.

Real Field

The purpose of the discussion in the previous section has been to show that $\mathbb{Q}$ has certain gaps. The real number system $\mathbb{R}$, which we will construct in this section, fills these gaps. This is the principal reason for the fundamental role which it plays in analysis.

The sole objective of this subsection is to prove the following result.

Theorem 14.20 (Existence of real field)**.** *There exists an ordered field $\mathbb{R}$ that*

(i) contains $\mathbb{Q}$ as a subfield, and

(ii) has the least-upper-bound property.

Remark. (ii) is also known as the *completeness axiom*.

We want to construct $\mathbb{R}$ from $\mathbb{Q}$; one method to do so is using Dedekind cuts.

Definition 14.21 (Dedekind cut)**.** $\alpha \subset \mathbb{Q}$ is a ***Dedekind cut***, if

(i) $\alpha \neq \emptyset$, $\alpha \neq \mathbb{Q}$; (non-trivial)

(ii) if $p \in \alpha$, $q \in \mathbb{Q}$ and $q < p$, then $q \in \alpha$;

(iii) if $p \in \alpha$, then $p < r$ for some $r \in \alpha$.

Remark. Note that (iii) simply says that α has no largest member; (ii) implies two facts which will be used freely:

- If $p \in \alpha$ and $q \notin \alpha$, then $p < q$.

- If $r \notin \alpha$ and $r < s$, then $s \notin \alpha$.

Example. Let $r \in \mathbb{Q}$ and define

$$\alpha_r := \{p \in \mathbb{Q} \mid p < r\}.$$

We now check that this is indeed a Dedekind cut.

(i) $p = 1 + r \notin \alpha_r$ thus $\alpha_r \neq \mathbb{Q}$. $p = r - 1 \in \alpha_r$ thus $\alpha_r \neq \emptyset$.

(ii) Suppose that $q \in \alpha_r$ and $q' < q$. Then $q' < q < r$ which implies that $q' < r$ thus $q' \in \alpha_r$.

(iii) Suppose that $q \in \alpha_r$. Consider $\frac{q+r}{2} \in \mathbb{Q}$ and $q < \frac{q+r}{2} < r$. Thus $\frac{q+r}{2} \in \alpha_r$.

This example shows that every rational r corresponds to a Dedekind cut α_r.

Example. $\sqrt[3]{2}$ is not rational, but it is real. $\sqrt[3]{2}$ corresponds to the cut

$$\alpha = \{p \in \mathbb{Q} \mid p^3 < 2\}.$$

(i) Trivial.

(ii) If $q < p$, by the monotonicity of the cubic function, this implies that $q^3 < p^3 < 2$ thus $q \in \alpha$.

(iii) If $p \in \alpha$, consider $\left(p + \frac{1}{n}\right)^3 < 2$.

Definition 14.22. The set of real numbers, denoted by $\mathbb{R}$, is the set of all Dedekind cuts:

$$\mathbb{R} := \{\alpha \subset \mathbb{Q} \mid \alpha \text{ is a Dedekind cut}\}.$$

Proposition 14.23. *$\mathbb{R}$ has an order, where $\alpha < \beta$ is defined to mean that $\alpha \subsetneq \beta$.*

Proof. Simply check if this is a valid order (by checking for trichotomy and transitivity). □

Proposition 14.24. *The ordered set $\mathbb{R}$ has the least-upper-bound property.*

Proof. Let non-empty $A \subset \mathbb{R}$ be bounded above. Let $\beta \in \mathbb{R}$ be an upper bound of A. We want to show that A has a supremum in $\mathbb{R}$.

Let

$$\gamma = \bigcup_{\alpha \in A} \alpha.$$

Then $p \in \gamma$ if and only if $p \in \alpha$ for some $\alpha \in A$.

Claim. $\gamma \in \mathbb{R}$ and $\gamma = \sup A$.

We first prove that $\gamma \in \mathbb{R}$ by checking that it is a Dedekind cut:

(i) Since $A \neq \emptyset$, there exists $\alpha_0 \in A$. Since $\alpha_0 \in \mathbb{R}$, it is a Dedekind cut so $\alpha_0 \neq \emptyset$. Since $\alpha_0 \subset \gamma$, $\gamma \neq \emptyset$.

Since $\alpha \subset \beta$ for every $\alpha \in A$, the union of $\alpha \in A$ must be a subset of β; thus $\gamma \subset \beta$. Hence $\gamma \neq \mathbb{Q}$.

(ii) Let $p \in \gamma$. Then $p \in \alpha_1$ for some $\alpha_1 \in A$. If $q < p$, then $q \in \alpha_1$ (since α_1 is a Dedekind cut). Hence $q \in \gamma$.

(iii) If $r \in \alpha_1$ is so chosen that $r > p$, we see that $r \in \gamma$ (since $\alpha_1 \subset \gamma$).

Next we prove that $\gamma = \sup A$, by checking that (i) γ is an upper bound of A, (ii) γ is the *least* of the upper bounds.

(i) It is clear that $\alpha \leq \gamma$ for every $\alpha \in A$.

(ii) Suppose $\delta < \gamma$. Then there exists $s \in \gamma$ such that $s \notin \delta$. Since $s \in \gamma$, $s \in \alpha$ for some $\alpha \in A$. Hence $\delta < \alpha$, so δ is not an upper bound of A.

□

Remark. The l.u.b. property of $\mathbb{R}$ is also known as the *completeness axiom* of $\mathbb{R}$.

We now define operations on $\mathbb{R}$.

Definition 14.25 (Addition)**.** Given $\alpha, \beta \in \mathbb{R}$, define addition as

$$\alpha + \beta := \{r \in \mathbb{Q} \mid r = a + b, a \in \alpha, b \in \beta\}.$$

We check that addition is closed in $\mathbb{R}$: for all $\alpha, \beta \in \mathbb{R}$, $\alpha + \beta \in \mathbb{R}$.

Proof. We check that $\alpha + \beta$ is a Dedekind cut:

(i) Since $\alpha \neq \emptyset$ and $\beta \neq \emptyset$, there exists $a \in \alpha$ and $b \in \beta$. Hence $r = a + b \in \alpha + \beta$ so $\alpha + \beta \neq \emptyset$.

Since $\alpha \neq \mathbb{Q}$ and $\beta \neq \mathbb{Q}$, there exist $c \neq \alpha$ and $d \neq \beta$. Thus $r' = c + d > a + b$ for any $a \in \alpha, b \in \beta$, so $r' \notin \alpha + \beta$. Hence $\alpha + \beta \neq \mathbb{Q}$.

(ii) Suppose that $r \in \alpha + \beta$ and $r' < r$. We want to show that $r' \in \alpha + \beta$.

$r = a + b$ for some $a \in \alpha, b \in \beta$. Then $r' - a < b$. Since $\beta \in \mathbb{R}$, $r' - a \in \beta$ so $r' - a = b_1$ for some $b_1 \in \beta$. Hence $r' = a + b_1 \in \alpha + \beta$.

(iii) Suppose $r \in \alpha+\beta$, so $r = a+b$ for some $a \in \alpha, b \in \beta$. Since α, β are Dedekind cuts, there exist $a' \in \alpha, b' \in \beta$ with $a < a'$ and $b < b'$. Then $r = a+b < a'+b' \in \alpha+\beta$. We define $r' = a'+b' \in \alpha+\beta$ with $r < r'$.

□

Lemma 14.26.

(i) Addition on $\mathbb{R}$ is commutative: $\alpha+\beta = \beta+\alpha$ for all $\alpha, \beta \in \mathbb{R}$.

(ii) Addition on $\mathbb{R}$ is associative: $\alpha+(\beta+\gamma) = (\alpha+\beta)+\gamma$ for all $\alpha, \beta, \gamma \in \mathbb{R}$.

(iii) Additive identity: Define $0^* := \{p \in \mathbb{Q} \mid p < 0\}$. Then $\alpha + 0^* = \alpha$ for all $\alpha \in \mathbb{R}$.

(iv) Additive inverse: Fix $\alpha \in \mathbb{R}$, define $\beta = \{p \in \mathbb{Q} \mid \exists r > 0, -p-r \notin \alpha\}$. Then $\alpha+\beta = 0^*$.

Remark. Recall that to prove that two sets are equal, show double inclusion.

Proof.

(i) We need to show that $\alpha+\beta \subset \beta+\alpha$ and $\beta+\alpha \subset \alpha+\beta$.

Let $r \in \alpha+\beta$. Then $r = a+b$ for $a \in \alpha$ and $b \in \beta$. Thus $r = b+a$ since $+$ is commutative on $\mathbb{Q}$. Hence $r \in \beta+\alpha$. Therefore $\alpha+\beta \subset \beta+\alpha$.

Similarly, $\beta+\alpha \subset \alpha+\beta$.

Therefore $\alpha+\beta = \beta+\alpha$.

(ii) Let $r \in \alpha+(\beta+\gamma)$. Then $r = a+(b+c)$ where $a \in \alpha, b \in \beta, c \in \gamma$. Thus $r = (a+b)+c$ by associativity of $+$ on $\mathbb{Q}$. Therefore $r \in (\alpha+\beta)+\gamma$, hence $\alpha+(\beta+\gamma) \subset (\alpha+\beta)+\gamma$.

Similarly, $(\alpha+\beta)+\gamma \subset \alpha+(\beta+\gamma)$.

(iii) It is clear that 0^* is a Dedekind cut.

Let $r \in \alpha+0^*$. Then $r = a+p$ for some $a \in \alpha, p \in 0^*$. Thus $r = a+p < a+0 = a$ so $r \in \alpha$. Hence $\alpha+0^* \subset \alpha$.

Let $r \in \alpha$. Then there exists $r' \in \alpha$ where $r' > r$. Thus $r-r' < 0$, so $r-r' \in 0^*$. We see that $r = r'+(r-r')$ where $r' \in \alpha$, $r-r' \in 0^*$. Hence $\alpha \subset \alpha+0^*$.

(iv) Fix some $\alpha \in \mathbb{R}$. We first show that β is a Dedekind cut.

(i) Let $s \notin \alpha$, let $p = -s-1$. Then $-p-1 \notin \alpha$. Hence $p \in \beta$, so $\beta \neq \emptyset$.
Let $q \in \alpha$. Then $-q \notin \beta$ so $\beta \neq \mathbb{Q}$.

(ii) Let $p \in \beta$. Then there exists $r > 0$ such that $-p-r \notin \alpha$. If $q < p$, then $-q-r > -p-r$ so $-q-r \notin \alpha$. Hence $q \in \beta$.

(iii) Let $t = p + \frac{r}{2}$. Then $t > p$, and $-t - \frac{r}{2} = -p - r \notin \alpha$. Hence $t \in \beta$.

Let $r \in \alpha$, $s \in \beta$. Then $-s \notin \alpha$. This implies $r < -s$ (since α is closed downwards) so $r + s < 0$. Hence $\alpha + \beta \subset 0^*$.

To prove the opposite inclusion, let $v \in 0^*$, and let $w = -\frac{v}{2}$. Then $w > 0$. By the Archimedean property on $\mathbb{Q}$, there exists $n \in \mathbb{N}$ such that $nw \in \alpha$ but $(n+1)w \notin \alpha$. Let $p = -(n+2)w$. Then

$$-p - w = (n+2)w - w = (n+1)w \notin \alpha$$

so $p \in \beta$. Since $v = nw + p$ where $nw \in \alpha$, $p \in \beta$, $v \in \alpha + \beta$. Hence $0^* \subset \alpha + \beta$.

□

Notation. β is denoted by the more familiar notation $-\alpha$.

Lemma 14.27. *If $\alpha, \beta, \gamma \in \mathbb{R}$ and $\beta < \gamma$, then $\alpha + \beta < \alpha + \gamma$.*

Proof. □

We say that a Dedekind cut α is *positive* if $0 \in \alpha$, and *negative* if $0 \notin \alpha$. If α is neither positive nor negative, then $\alpha = 0^*$.

Multiplication is a little more bothersome than addition in the present context, since products of negative rationals are positive. For this reason we confine ourselves first to $\mathbb{R}^+$ (the set of all $\alpha \in \mathbb{R}$ with $\alpha > 0^*$).

Definition 14.28. Given $\alpha, \beta \in \mathbb{R}^+$, define multiplication as

$$\alpha\beta := \{p \in \mathbb{Q} \mid p \leq rs,\ r \in \alpha, s \in \beta,\ r, s > 0\}.$$

We also define $1^* := \{q \in \mathbb{Q} \mid q < 1\}$.

We check that multiplication is closed in $\mathbb{R}^+$: for all $\alpha, \beta \in \mathbb{R}^+$, $\alpha\beta \in \mathbb{R}^+$.

Proof. Check that $\alpha\beta$ is a Dedekind cut.

(i) $\alpha \neq \emptyset$ means there exists $r \in \alpha, r > 0$. Similarly, $\beta \neq \emptyset$ means there exists $s \in \beta, s > 0$. Then $rs \in \mathbb{Q}$ and $rs \leq rs$, so $rs \in \alpha\beta$. Hence $\alpha\beta \neq \emptyset$.

$\alpha \neq \mathbb{Q}$ means there exists $r' \notin \alpha$ such that $r' > r$ for all $r \in \alpha$. Similarly $\beta \neq \mathbb{Q}$ means there exists $s' \in \beta$ such that $s' > s$ for all $s \in \beta$. Then $r's' > rs$ for all $r \in \alpha, s \in \beta$, so $r's' \notin \alpha\beta$. Hence $\alpha\beta \neq \mathbb{Q}$.

(ii) Let $p \in \alpha\beta$. Then $p \leq ab$ for some $a \in \alpha, b \in \beta$, $a, b > 0$.

If $q < p$, then $q < p \leq ab$ so $q \in \alpha\beta$.

(iii) Let $p \in \alpha\beta$. Then $p \leq ab$ for some $a \in \alpha, b \in \beta$, $a, b > 0$. Pick $a' \in \alpha$ and $b' \in \beta$ with $a' > a$ and $b' > b$. Form $a'b' > ab \geq p$, $a'b' \leq a'b'$ means $a'b' \in \alpha \cdot \beta$.

□

We now complete the definition of multiplication by setting $\alpha 0^* = 0^* = 0^* \alpha$, and by setting

$$\alpha\beta = \begin{cases} (-\alpha)(-\beta) & a < 0^*, \beta < 0^*, \\ -[(-\alpha)\beta] & a < 0^*, \beta > 0^*, \\ -[\alpha(-\beta)] & \alpha > 0^*, \beta < 0^*. \end{cases}$$

where we make negative numbers positive, multiply, and then negate them as needed.

Lemma 14.29.

(i) Multiplication on $\mathbb{R}$ is commutative: $\alpha\beta = \beta\alpha$ for all $\alpha, \beta \in \mathbb{R}$.

(ii) Multiplication on $\mathbb{R}$ is associative: $(\alpha\beta)\gamma = \alpha(\beta\gamma)$ for all $\alpha, \beta, \gamma \in \mathbb{R}$.

(iii) Multiplicative identity: $1\alpha = \alpha$ for all $\alpha \in \mathbb{R}$.

(iv) Multiplicative inverse: If $\alpha \in \mathbb{R}$, $\alpha \neq 0^*$, then there exists $\beta \in \mathbb{R}$ such that $\alpha\beta = 1^*$.

We associate each $r \in \mathbb{Q}$ with the set

$$r^* = \{p \in \mathbb{Q} \mid p < r\}.$$

It is obvious that each r^* is a cut; that is, $r^* \in \mathbb{R}$.

Proposition 14.30. *The replacement of $r \in \mathbb{Q}$ by the corresponding "rational cuts" $r^* \in \mathbb{R}$ preserves sums, products, and order. That is, for all $r^*, s^* \in \mathbb{R}$,*

(i) $r^* + s^* = (r+s)^*$;

(ii) $r^* s^* = (rs)^*$;

(iii) $r^* < s^*$ if and only if $r < s$.

Proof.

(i) Let $p \in r^* + s^*$. Then $p = u + v$ for some $u \in r^*$, $v \in s^*$, where $u < r$, $v < s$. Then $p < r + s$. Hence $p \in (r+s)^*$, so $r^* + s^* \subset (r+s)^*$.

Let $p \in (r+s)^*$. Then $p < r+s$. Let $t = \dfrac{(r+s) - p}{2}$, and let

$$r' = r - t, \quad s' = s - t.$$

Since $t > 0$, $r' < r$ so $r' \in r^*$; $s' < s$ so $s' \in s^*$. Then $p = r' + s'$, so $p \in r^* + s^*$. Hence $(r+s)^* \subset r^* + s^*$.

(ii)

(iii) Suppose $r < s$. Then $r \in s^*$, but $r \notin r^*$. Hence $r^* < s^*$.

Conversely, suppose $r^* < s^*$. Then there exists $p \in s^*$ such that $p \in r^*$. Hence $r \leq p < s$, so $r < s$.

□

This shows that the ordered field $\mathbb{Q}$ is isomorphic to the ordered field $\mathbb{Q}^* = \{q^* \mid q \in \mathbb{Q}\}$ whose elements are rational cuts. It is this identification of $\mathbb{Q}$ with $\mathbb{Q}^*$ which allows us to regard $\mathbb{Q}$ as a subfield of $\mathbb{R}$.

Remark. In fact, $\mathbb{R}$ is the only ordered field with the l.u.b. property. Hence any other ordered field with the l.u.b. property is isomorphic to $\mathbb{R}$.

Therefore we have proven 14.20.

Properties of $\mathbb{R}$

Proposition 14.31 (Archimedean property). *For any $x \in \mathbb{R}^+$, $y \in \mathbb{R}$, there exists $n \in \mathbb{N}$ such that*

$$nx > y.$$

Proof. Suppose, for a contradiction, that $nx \leq y$ for all $n \in \mathbb{N}$. Then y is an upper bound of the set

$$A = \{nx \mid n \in \mathbb{N}\}.$$

Since $A \subset \mathbb{R}$ is non-empty and bounded above, by the l.u.b. property of $\mathbb{R}$, A has a supremum in $\mathbb{R}$, say $\alpha = \sup A$.

Consider $\alpha - x$. Since $\alpha - x < \alpha = \sup A$, $\alpha - x$ is not an upper bound of A. Thus $\alpha - x \leq mx$ for some $m \in \mathbb{N}$; rearranging gives $\alpha \leq (m+1)x$. This implies α is not an upper bound of A, a contradiction. □

Corollary 14.32. *Let $\varepsilon > 0$. Then there exists $n \in \mathbb{N}$ such that $0 < \frac{1}{n} < \varepsilon$.*

Proof. In 14.31, take $x = \varepsilon$ and $y = 1$. □

A subset $D \subset \mathbb{R}$ is said to be *dense* in $\mathbb{R}$ if for any $a, b \in \mathbb{R}$ with $a < b$, $D \cap (a,b) \neq \emptyset$.

Proposition 14.33 ($\mathbb{Q}$ is dense in $\mathbb{R}$). *For any $x, y \in \mathbb{R}$ with $x < y$, there exists $p \in \mathbb{Q}$ such that*

$$x < p < y.$$

Proof. We prove by construction (construct the required p from the given x and y).

Since $x < y$, we have $y - x > 0$. By 14.32, there exists $n \in \mathbb{N}$ such that

$$\frac{1}{n} < y - x.$$

Consider $E = \left\{ \frac{k}{n} \mid k \in \mathbb{N} \right\}$, the set of multiples of $\frac{1}{n}$. Since E is unbounded, choose the smallest $m \in \mathbb{N}$ such that $\frac{m}{n} > x$.

Claim. $x < \frac{m}{n} < y$.

We only need to show $\frac{m}{n} < y$. Suppose, for a contradiction, that this does not hold. Then

$$\frac{m-1}{n} < x \quad \text{and} \quad \frac{m}{n} > y,$$

where the first inequality follows from the minimality of m. But these two statements combined imply that $\frac{1}{n} > y - x$, a contradiction. □

Corollary 14.34 ($\mathbb{R} \setminus \mathbb{Q}$ is dense in $\mathbb{R}$). *For any $x, y \in \mathbb{R}$ with $x < y$, then there exists an irrational number r such that $x < r < y$.*

Proof. By the density theorem, there exists $p \in \mathbb{Q}$ such that $p \neq 0$ and $\frac{x}{\sqrt{2}} < p < \frac{y}{\sqrt{2}}$. Thus

$$x < p\sqrt{2} < y$$

and $p\sqrt{2}$ is irrational. □

Corollary 14.35. *Every interval $I \subset \mathbb{R}$ contains infinitely many rational numbers and infinitely many irrational numbers.*

Proposition 14.36 ($\mathbb{R}$ is closed under taking roots). *For every $x \in \mathbb{R}^+$ and every $n \in \mathbb{N}$, there exists a unique $y \in \mathbb{R}^+$ so that $y^n = x$.*

We call the number y the positive n-th root of x, and denote it by $\sqrt[n]{x}$ or $x^{1/n}$.

Proof. Let $x \in \mathbb{R}^+$, fix $n \in \mathbb{N}$.

Existence Let

$$E = \{t \in \mathbb{R}^+ \mid t^n < x\}.$$

Claim. $y = \sup E$ satisfies $y^n = x$.

We first show that E has a supremum.

(i) Let $t = \dfrac{x}{1+x}$. Then $0 \le t < 1$, so $t^n \le t < x$ implies $t^n < x$. Hence $t \in E$, so $E \neq \emptyset$.

(ii) We claim that $1+x$ is an upper bound for E.

If $t > 1+x$, then $t^n \ge t > x$ implies $t^n > x$, so $t \notin E$. [This is the contrapositive of $t \in E \implies t \le 1+x$.] Hence $1+x$ is an upper bound of E, so E is bounded above.

Hence E has a supremum; let $y = \sup E$.

To prove $y^n = x$, we show both $y^n < x$ and $y^n > x$ lead to a contradiction. Consider the identity $b^n - a^n = (b-a)\left(b^{n-1} + b^{n-2}a + \cdots + a^{n-1}\right)$. If $0 < a < b$, then

$$b^n - a^n < (b-a)nb^{n-1}. \tag{1}$$

Case 1: $y^n < x$**.**

Idea. We can find a *small* $h > 0$ such that $(y+h)^n < x$.

Choose h so that $0 < h < 1$ and

$$h < \frac{x - y^n}{n(y+1)^{n-1}}.$$

In (1), take $b = y+h$, $a = y$. Then

$$\begin{aligned}(y+h)^n - y^n &< hn(y+h)^{n-1} \\ &< hn(y+1)^{n-1} \\ &< \frac{x-y^n}{\cancel{n(y+1)^{n-1}}}\cancel{n(y+1)^{n-1}} \\ &= x - y^n.\end{aligned}$$

Thus $(y+h)^n < x$, and $y+h \in E$. Since $y+h > y$, this contradicts the fact that y is an upper bound of E.

Case 2: $y^n > x$**.**

Idea. Similarly, we can find a *small* $k > 0$ such that $(y-k)^n > x$.

Let

$$k = \frac{y^n - x}{ny^{n-1}}.$$

Then $0 < k < y$, by (1). If $t \geq y - k$,

$$\begin{aligned} y^n - t^n &\leq y^n - (y-k)^n \\ &< kny^{n-1} \\ &= \frac{y^n - x}{\cancel{ny^{n-1}}} \cancel{ny^{n-1}} \\ &= y^n - x. \end{aligned}$$

Thus $t^n > x$, and $t \notin E$. It follows that $y - k$ is an upper bound of E. But $y - k < y$, which contradicts the fact that y is the *least* upper bound of E.

Uniqueness Suppose, for a contradiction, that there exist distinct y_1, y_2 which are both n-th roots of x. WLOG assume that $0 < y_1 < y_2$. Then taking the n-th power gives ${y_1}^n < {y_2}^n$.

Since y_1 is a n-th root of x, then $x = {y_1}^n$, so $x < {y_2}^n$ implies $x \neq {y_2}^n$. Hence y_2 cannot be a n-th root of x, a contradiction. □

Corollary 14.37. *If $a, b \in \mathbb{R}^+$ and $n \in \mathbb{N}$, then*

$$(ab)^{\frac{1}{n}} = a^{\frac{1}{n}} b^{\frac{1}{n}}.$$

Proof. Let $\alpha = a^{1/n}$, $\beta = b^{1/n}$. Then

$$ab = \alpha^n \beta^n = (\alpha\beta)^n$$

where the last line follows from commutativity of multiplication. The uniqueness assertion of the previous result allows us to take the n-th root on both sides:

$$(ab)^{\frac{1}{n}} = \alpha\beta = a^{\frac{1}{n}} b^{\frac{1}{n}}.$$

□

Lemma 14.38. *If $a \in \mathbb{R}^+$ and $m, n \in \mathbb{N}$, then*

$$(a^{1/n})^m = (a^m)^{1/n}.$$

Proof. We have

$$\left(\left(a^{1/n}\right)^m\right)^n = \left(a^{1/n}\right)^{mn} = \left(\left(a^{1/n}\right)^n\right)^m = a^m.$$

By definition, this yields the desired result. □

For $a \in \mathbb{R}^+$ and $m, n \in \mathbb{N}$, we define *rational exponents*

$$a^{m/n} := \left(a^{1/n}\right)^m \quad \text{and} \quad a^{-m/n} := \frac{1}{a^{m/n}}.$$

(We also define $a^0 = 1$.)

We need to check that the above definition of a^r is well defined. That is, if $m,n,p,q \in \mathbb{N}$ are such that $\frac{m}{n} = \frac{p}{q}$, then $\left(a^{1/n}\right)^m = \left(a^{1/q}\right)^p$. To see this, note that $mq = np$, and thus

$$\left(\left(a^{1/n}\right)^m\right)^q = \left(a^{1/n}\right)^{mq} = \left(a^{1/n}\right)^{np} = a^p.$$

Hence $\left(a^{1/n}\right)^m$ is the q-th root of a^p, i.e.,

$$(a^{1/n})^m = (a^p)^{1/q}.$$

Lemma 14.39 (Properties of rational exponents)**.**

(i) If $a > 0$ and $r,s \in \mathbb{Q}$, then $a^{r+s} = a^r a^s$ and $(a^r)^s = a^{rs}$.

(ii) If $0 < a < b$ and $r \in \mathbb{Q}$ with $r > 0$, then $a^r < b^r$.

(iii) If $a > 1$, $r,s \in \mathbb{Q}$ with $r < s$, then $a^r < a^s$.

The next result shows that real numbers can be approximated to any desired degree of accuracy by rational numbers with finite decimal representations.

Proposition 14.40. *Let $x \geq 0$. Then for every integer $n \geq 1$ there exists a finite decimal $r_n = a_0.a_1a_2\cdots a_n$ such that*

$$r_n \leq x < r_n + \frac{1}{10^n}.$$

Proof. We prove by construction (construct the required finite decimal from x).

Let

$$S = \{k \in \mathbb{Z} \mid k \leq x\}.$$

S is non-empty (since $0 \in S$), and S is bounded above by x. Hence by the lub property of $\mathbb{R}$, S has a supremum in $\mathbb{R}$, say $a_0 = \sup S$. It is easily verified that $a_0 \in S$, so a_0 is a non-negative integer. We call a_0 the *greatest integer* in x, and write $a_0 = \lfloor x \rfloor$. Clearly we have

$$a_0 \leq x < a_0 + 1.$$

Now let $a_1 = \lfloor 10(x - a_0) \rfloor$. Since $0 \leq 10(x - a_0) < 10$, we have $0 \leq a_1 \leq 9$ and

$$a_1 \leq 10x - 10a_0 < a_1 + 1.$$

In other words, a_1 is the largest integer satisfying the inequalities

$$a_0 + \frac{a_1}{10} \leq x < a_0 + \frac{a_1 + 1}{10}.$$

More generally, having chosen $a_1, \ldots, a_{n-1}$ with $0 \le a_i \le 9$, let a_n be the largest integer satisfying the inequalities

$$a_0 + \frac{a_1}{10} + \cdots + \frac{a_n}{10^n} \le a_0 + \frac{a_1}{10} + \cdots + \frac{a_n + 1}{10^n}.$$

Then $0 \le a_n \le 9$ and we have

$$r_n \le x < r_n + \frac{1}{10^n},$$

where $r_n = a_0.a_1a_2\cdots a_n$. □

Furthermore, it is easy to verify that $x = \sup_{n\in\mathbb{N}} r_n$.

Extended Real Number System

Definition 14.41. Define the ***extended real number system*** as $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$.

Notation. We sometimes write $[-\infty, \infty]$ in place of $\overline{\mathbb{R}}$.

We preserve the original order in $\mathbb{R}$, and define

$$-\infty < x < +\infty$$

for all $x \in \mathbb{R}$.

Defining $\overline{\mathbb{R}}$ is convenient since the following result holds.

Lemma 14.42. *Any non-empty $E \subset \overline{\mathbb{R}}$ has a supremum and infimum in $\overline{\mathbb{R}}$.*

Proof. If E is bounded above in $\mathbb{R}$, then by the l.u.b. property of $\mathbb{R}$, it has a supremum in $\mathbb{R} \subset \overline{\mathbb{R}}$. If E is not bounded above in $\mathbb{R}$, then $\sup E = +\infty \in \overline{\mathbb{R}}$.

Exactly the same remarks apply to lower bounds. □

$\overline{\mathbb{R}}$ does not form a field, but it is customary to make the following conventions for arithmetic on $\overline{\mathbb{R}}$:

(i) If x is real or ∞, then $x + \infty = +\infty$.

If x is real or $-\infty$, then $x - \infty = -\infty$.

(ii) If $x > 0$, then $x \cdot (+\infty) = +\infty, \quad x \cdot (-\infty) = -\infty$.

If $x < 0$, then $x \cdot (+\infty) = -\infty, \quad x \cdot (-\infty) = +\infty$.

If x is real, then $\dfrac{x}{+\infty} = \dfrac{x}{-\infty} = 0$.

(Note that addition and multiplication are understood to be commutative on the extended reals, so that the definitions also imply further cases such as $+\infty + x = +\infty$.)

14.3 Complex Field

Lemma 14.43. *Let $(a,b),(c,d) \in \mathbb{R}^2$. Define addition and multiplication on $\mathbb{R}^2$ as*

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

Then $\mathbb{R}^2$ is a field, with additive identity $(0,0)$ and multiplicative identity $(1,0)$.

We call this structure $\mathbb{C}$, the ***complex field***; its elements are called *complex numbers*.

Proof. Check the field axioms. □

The next result shows that the complex numbers of the form $(a,0)$ have the same arithmetic properties as the corresponding real numbers a. We can therefore identify $(a,0) \in \mathbb{C}$ with $a \in \mathbb{R}$. This identification implies that $\mathbb{R}$ is a *subfield* of $\mathbb{C}$.

Lemma 14.44. *For any $a,b \in \mathbb{R}$,*

$$(a,0)+(b,0) = (a+b,0),$$
$$(a,0)(b,0) = (ab,0).$$

Proof. Exercise. □

You may have noticed that we have defined the complex numbers without referring to the mysterious square root of -1. We now show that the notation (a,b) is equivalent to the more customary $a+bi$.

Define the imaginary number $i := (0,1)$. See that

$$i^2 = (0,1)(0,1) = (-1,0) = -1.$$

Lemma 14.45. *For any $a,b \in \mathbb{R}$, $(a,b) = a+bi$.*

Proof. $a+bi = (a,0)+(b,0)(0,1) = (a,0)+(0,b) = (a,b)$. □

For $a,b \in \mathbb{R}$, we write $z = a+bi$; we call a and b the *real part* and *imaginary part* of z respectively, denoted by $a = \mathrm{Re}(z)$, $b = \mathrm{Im}(z)$. We call $\bar{z} = a-bi$ the ***conjugate*** of z.

The next result summarises basic properties of the conjugate of a complex number.

Lemma 14.46. *For $z, w \in \mathbb{C}$,*

(i) $\overline{z+w} = \overline{z} + \overline{w}$

(ii) $\overline{zw} = \overline{z}\,\overline{w}$

(iii) $z + \overline{z} = 2\,\mathrm{Re}(z)$, $z - \overline{z} = 2i\,\mathrm{Im}(z)$

(iv) $z\overline{z}$ *is real, and* $z\overline{z} \geq 0$

Proof. Let $z = a + bi$, $w = c + di$.

(i) $\overline{z+w} = \overline{(a+c)+(b+d)i} = (a+c) - (b+d)i = (a-bi) + (c-di) = \overline{z} + \overline{w}$.

(ii) $\overline{zw} = \overline{(a+bi)(c+di)} = \overline{(ac-bd)+(ad+bc)i} = (ac-bd) - (ad+bc)i = (a-bi)(c-di) = \overline{z}\,\overline{w}$.

(iii) $z + \overline{z} = (a+bi) + (a-bi) = 2a = 2\,\mathrm{Re}(z)$.
$z - \overline{z} = (a+bi) - (a-bi) = 2bi = 2i\,\mathrm{Im}(z)$.

(iv) $z\overline{z} = (a+bi)(a-bi) = a^2 + b^2 \in \mathbb{R}_{\geq 0}$.

□

Let $z \in \mathbb{C}$. The ***absolute value*** of z is defined as

$$|z| := (z\overline{z})^{1/2}.$$

The existence (and uniqueness) of $|z|$ follows from 14.36, and (iv) of the previous result. Note that when z is real, then $\overline{z} = z$; thus $|z| = \sqrt{z^2}$. Hence $|z| = z$ if $z \geq 0$, and $|z| = -z$ if $z < 0$.

Remark. Since the absolute value is defined as a square root, it is more useful to work with the *square* of the absolute value.

The next result summarises basic properties of the absolute value.

Lemma 14.47. *For $z, w \in \mathbb{C}$,*

(i) $|z| \geq 0$

(ii) $|\overline{z}| = |z|$

(iii) $|zw| = |z||w|$

(iv) $|\mathrm{Re}(z)| \leq |z|$

Proof.

(i) The square root is non-negative, by definition.

(ii) The conjugate of $\bar{z}$ is z, and the rest follows by the definition of absolute value.

(iii) Let $z = a+bi$, $w = c+di$ where $a,b,c,d \in \mathbb{R}$. Then

$$\begin{aligned} |zw|^2 &= (ac-bd)^2 + (ad-bc)^2 \\ &= (a^2+b^2)(c^2+d^2) \\ &= |z|^2|w|^2 = (|z||w|)^2 \end{aligned}$$

and the desired result follows by taking square roots on both sides.

(iv) Let $z = a+bi$. Note that $a^2 \leq a^2+b^2$, hence

$$|\operatorname{Re}(z)| = |a| = \sqrt{a^2} \leq \sqrt{a^2+b^2} = |z|.$$

□

Proposition 14.48 (Triangle inequality). *For $z,w \in \mathbb{C}$,*

$$|z+w| \leq |z|+|w|. \tag{14.1}$$

Proof. Let $z,w \in \mathbb{C}$. Note that the conjugate of $z\overline{w}$ is $\bar{z}w$, so $z\overline{w}+\bar{z}w = 2\operatorname{Re}(z\overline{w})$. Hence

$$\begin{aligned} |z+w|^2 &= (z+w)(\overline{z+w}) = (z+w)(\bar{z}+\overline{w}) \\ &= z\bar{z}+z\overline{w}+\bar{z}w+w\overline{w} \\ &= |z|^2+2\operatorname{Re}(z\overline{w})+|w|^2 \\ &\leq |z|^2+2|z\overline{w}|+|w|^2 \\ &= |z|^2+2|z||w|+|w|^2 \\ &= (|z|+|w|)^2 \end{aligned}$$

and taking square roots yields the desired result. □

Corollary 14.49 (Generalised triangle inequality). *For $z_1,\ldots,z_n \in \mathbb{C}$,*

$$|z_1+\cdots+z_n| \leq |z_1|+\cdots+|z_n|.$$

Proof. Induct on n. The case when $n=1$ is trivial. We have proven the case when $n=2$. Assume the statement holds for $n-1$. Then

$$|z_1+\cdots+z_{n-1}+z_n| \leq |z_1+\cdots+z_{n-1}|+|z_n| \leq |z_1|+\cdots+|z_n|.$$

□

Corollary 14.50. *For $x, y, z \in \mathbb{C}$,*

(i) $||x| - |y|| \leq |x - y|$;

(ii) $|x - y| \leq |x - z| + |z - y|$.

Proof.

(i) By the triangle inequality,

$$|x| = |(x - y) + y| \leq |x - y| + |y|$$

so that

$$|x| - |y| \leq |x - y|.$$

Interchanging the roles of x and y in the above gives

$$|y| - |x| \leq |x - y|.$$

Hence

$$||x| - |y|| \leq |x - y|.$$

(ii) In the triangle inequality, replace x by $x - y$ and y by $y - z$.

□

Proposition 14.51 (Cauchy–Schwarz inequality)**.** *If $a_1, \ldots, a_n, b_1, \ldots, b_n \in \mathbb{C}$, then*

$$\left| \sum_{i=1}^{n} a_i \overline{b_i} \right|^2 \leq \sum_{i=1}^{n} |a_i|^2 \sum_{i=1}^{n} |b_i|^2. \tag{14.2}$$

Proof. For simplicity, we shall drop the upper and lower limits of the sums. Let

$$A = \sum |a_i|^2, \quad B = \sum |b_i|^2, \quad C = \sum a_i \overline{b_i}.$$

Then (14.2) becomes

$$|C|^2 \leq AB.$$

If $B = 0$, then $b_1 = \cdots = b_n = 0$, and the conclusion is trivial. Now assume that $B > 0$. Then

consider the sum

$$\begin{aligned}\sum |Ba_i - Cb_i|^2 &= \sum (Ba_i - Cb_i)(\overline{Ba_i - Cb_i}) \\ &= \sum (Ba_i - Cb_i)(B\overline{a_i} - \overline{Cb_i}) \\ &= B^2 \sum |a_i|^2 - B\overline{C} \sum a_i \overline{b_i} - BC \sum \overline{a_i} b_i + |C|^2 \sum |b_i|^2 \\ &= B^2 A - B|C|^2 \\ &= B(AB - |C|^2).\end{aligned}$$

Each term in $\sum |Ba_i - Cb_i|^2$ is non-negative, so $\sum |Ba_i - Cb_i|^2 \geq 0$. Thus

$$B(AB - |C|^2) \geq 0.$$

Since $B > 0$, it follows that $AB - |C|^2 \geq 0$, or $|C|^2 \leq AB$. This is the desired inequality.

(when does equality hold?) □

Define

$$\mathbb{C}^n = \{(z_1, \ldots, z_n) \mid z_i \in \mathbb{C}\}.$$

We can define an inner product on $\mathbb{C}^n$: for $\mathbf{a}, \mathbf{b} \in \mathbb{C}^n$,

$$\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i=1}^{n} a_i \overline{b_i}.$$

We can also define the norm of $\mathbf{a} \in \mathbb{C}^n$:

$$|\mathbf{a}| = \langle \mathbf{a}, \mathbf{a} \rangle^{\frac{1}{2}}.$$

14.4 Euclidean Space

For $n \in \mathbb{N}$, define

$$\mathbb{R}^n := \{(x_1, \ldots, x_n) \mid x_i \in \mathbb{R}\}$$

where $\mathbf{x} = (x_1, \ldots, x_n)$, x_i's are called the coordinates of $\mathbf{x}$. The elements of $\mathbb{R}^n$ are called *points*, or *vectors*.

Lemma 14.52. *Let $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{y} = (y_1, \ldots, y_n)$, $\alpha \in \mathbb{R}$. Define addition and scalar multiplication on $\mathbb{R}^n$ as*

$$\mathbf{x} + \mathbf{y} = (x_1 + y_1, \ldots, x_n + y_n),$$
$$\alpha\mathbf{x} = (\alpha x_1, \ldots, \alpha x_n).$$

Then $\mathbb{R}^n$ is a vector space over $\mathbb{R}$, with zero element $\mathbf{0} = (0, \ldots, 0)$.

Proof. These two operations satisfy the commutative, associatives, and distributive laws (the proof is trivial, in view of the analagous laws for the real numbers). □

We define the *inner product* of $\mathbf{x}$ and $\mathbf{y}$ by

$$\mathbf{x} \cdot \mathbf{y} := \sum_{i=1}^{n} x_i y_i,$$

and the *norm* of $\mathbf{x}$ by

$$\|\mathbf{x}\| := \sqrt{\mathbf{x} \cdot \mathbf{x}}.$$

The structure now defined (the vector space $\mathbb{R}^n$ with the above inner product and norm) is called the ***Euclidean n-space***.

The next result summarises basic properties of the norm on $\mathbb{R}^n$.

Lemma 14.53. *Suppose $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$.*

(i) $\|\mathbf{x}\| \geq 0$, *where equality holds if and only if* $\mathbf{x} = \mathbf{0}$ *(positive definiteness)*

(ii) $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$ *(homogeneity)*

(iii) $\|\mathbf{x} \cdot \mathbf{y}\| \leq \|\mathbf{x}\|\|\mathbf{y}\|$ *(Cauchy–Schwarz inequality)*

(iv) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ *(triangle inequality)*

(v) $\|\mathbf{x} - \mathbf{z}\| \leq \|\mathbf{x} - \mathbf{y}\| + \|\mathbf{y} - \mathbf{z}\|$ *(triangle inequality)*

Proof.

(i) Obvious from definition.

(ii) Obvious from definition.

(iii) We want to show

$$\sqrt{\sum_{i=1}^{n} x_i y_i} \leq \sqrt{\sum_{i=1}^{n} {x_i}^2} \sqrt{\sum_{i=1}^{n} {y_i}^2},$$

or, squaring both sides,

$$\sum_{i=1}^{n} x_i y_i \leq \left(\sum_{i=1}^{n} {x_i}^2 \right) \left(\sum_{i=1}^{n} {y_i}^2 \right).$$

But this is simply the Cauchy–Schwarz inequality (14.2).

(iv) By (iii) we have

$$\begin{aligned} \|\mathbf{x}+\mathbf{y}\| &= (\mathbf{x}+\mathbf{y}) \cdot (\mathbf{x}+\mathbf{y}) \\ &= \mathbf{x} \cdot \mathbf{x} + 2\mathbf{x} \cdot \mathbf{y} + \mathbf{y} \cdot \mathbf{y} \\ &\leq \|\mathbf{x}\|^2 + 2\|\mathbf{x}\|\|\mathbf{y}\| + \|\mathbf{y}\|^2 \\ &= (\|\mathbf{x}\| + \|\mathbf{y}\|)^2 . \end{aligned}$$

(v) This follows directly from (iv) by replacing $\mathbf{x}$ by $\mathbf{x}-\mathbf{y}$, and $\mathbf{y}$ by $\mathbf{y}-\mathbf{z}$.

□

Exercises

Exercise 14.1 ([Rud76] 1.1)**.** If $r \in \mathbb{Q} \setminus \{0\}$ and $x \in \mathbb{R} \setminus \mathbb{Q}$, prove that $r+x \in \mathbb{R} \setminus \mathbb{Q}$ and $rx \in \mathbb{R} \setminus \mathbb{Q}$.

Solution. Prove by contradiction. If r and $r+x$ were both rational, then $x = (r+x) - r$ would also be rational. Similarly if rx were rational, then $x = \frac{rx}{r}$ would also be rational. □

Exercise 14.2 ([Rud76] 1.2)**.** Prove that there is no rational number whose square is 12.

Solution. Prove by contradiction. □

Exercise 14.3 ([Rud76] 1.4)**.** Let E be a nonempty subset of an ordered set; suppose α is a lower bound of E, and β is an upper bound of E. Prove that $\alpha \leq \beta$.

Solution. Since E is non-empty, there exists $x \in E$. Since α is a lower bound of E, $\alpha \leq x$. Since β is an upper bound of E, $x \leq \beta$. Combining these two inequalities gives $\alpha \leq x \leq \beta$; thus $\alpha \leq \beta$. □

Exercise 14.4. *Finite sets always have suprema.*
Let S be an ordered set (not assumed to have the l.u.b. property).

(i) Show that every two-element subset $\{x, y\} \subset S$ has a supremum.

(ii) Deduce (using induction) that every finite subset of S has a supremum.

Solution.

(i) Use trichotomy: if $x \leq y$, the supremum is y; if $x > y$, the supremum is x.

(ii) We will show that for each $n \in \mathbb{N}$, every n-element subset of S has a supremum.

The case of a singleton is trivial. The case where $n = 2$ has been shown in (i).

Now suppose the desired result holds for n. Let $A = \{x_1, \ldots, x_n, x_{n+1}\} \subset S$. By induction hypothesis, $\{x_1, \ldots, x_n\}$ has a supremum x_k for some $k \in \{1, \ldots, n\}$. If $x_k \leq x_{n+1}$, $\sup A = x_{n+1}$; if $x_k > x_{n+1}$, $\sup A = x_k$.

□

Exercise 14.5. *If one set lies above another.*
Suppose S is a set with the l.u.b. property and the g.l.b. property, and suppose X and Y are non-empty subsets of S.

(i) If every element of X is $\leq$ every element of Y, show that $\sup X \leq \inf Y$.

(ii) If every element of X is $<$ every element of Y, does it follow that $\sup X < \inf Y$? (Give a proof or a counterexample.)

Solution.

(i) Since X is bounded above by elements of Y, let $\alpha = \sup X$. Similarly, since Y is bounded below by elements of X, let $\beta = \inf Y$.

Since every $y \in Y$ is an upper bound for X, and α is the *least* upper bound, we must have $\alpha \leq y$ for all $y \in Y$. Thus α is a lower bound of Y.

But β is the *greatest* lower bound of Y, so $\alpha \leq \beta$.

(ii) Consider $X = (0,1)$ and $Y = (1,2)$.

□

Exercise 14.6. *Least upper bounds of least upper bounds.*
Let S be an ordered set with the l.u.b. property, and let $\{A_i\}_{i\in I}$ be a non-empty family of non-empty subsets of S.

(i) Suppose each A_i is bounded above, let $\alpha_i = \sup A_i$, and suppose further that $\{\alpha_i\}_{i\in I}$ is bounded above. Show that $\bigcup_{i\in I} A_i$ is bounded above, and

$$\sup\left(\bigcup_{i\in I} A_i\right) = \sup_{i\in I} \alpha_i.$$

(ii) On the other hand, suppose either (a) not all of the A_i are bounded above, or (b) they are all bounded above, but writing $\alpha_i = \sup A_i$ for each i, the set $\{\alpha_i\}_{i\in I}$ is unbounded above. Show in each of these cases that $\bigcup_{i\in I} A_i$ is unbounded above.

(iii) Again suppose each A_i is bounded above, with $\alpha_i = \sup A_i$. Show that $\bigcap_{i\in I} A_i$ is also bounded above. Must it be non-empty? If it is non-empty, what can be said about the relationship between $\sup\left(\bigcap_{i\in I} A_i\right)$ and the numbers α_i $(i \in I)$.

Solution.

(i) Let $A = \bigcup_{i\in I} A_i$. We first show A is bounded above.

Let $\alpha = \sup_{i\in I} a_i$. Since $\alpha_i = \sup A_i$ is an upper bound of A_i, we have $a_i \leq \alpha_i$ for all $a_i \in A_i$. Hence for any $a \in A$, $a \in A_i$ for some A_i, so $a \leq \alpha_i \leq \alpha$. Thus A is bounded above, so $\sup A$ exists.

We have shown that α is an upper bound of A. We now show it is the *least* upper bound. Let $u < \alpha$. (We need to show that u is not an upper bound of A.)

Since $u < \alpha = \sup_{i\in I} \alpha_i$, by definition of supremum, there exists $i_0 \in I$ such that $\alpha_{i_0} > u$.

But $\alpha_{i_0} = \sup A_{i_0}$, so $u < \sup A_{i_0}$. Hence by definition of supremum, there exists $a \in A_{i_0}$ such that $a > u$.

Since $a \in A$ but $a > u$, we conclude that u is not an upper bound of A. Hence

$$\sup A = \alpha = \sup_{i\in I} \alpha_i.$$

(ii) Let $A = \bigcup_{i\in I} A_i$.

(a) Suppose A_{i_0} is not bounded above for some $i_0 \in I$. Then for any $u \in S$, there exists $a \in A_{i_0} \subset A$ such that $a > u$. Hence A is unbounded above.

(b) Let $u \in S$. Since $\{\alpha_i\}_{i\in I}$ is unbounded above, $\alpha_{i_0} > u$ for some $i_0 \in I$. But $\alpha_{i_0} = \sup A_{i_0}$, so by definition of supremum, there exists $a \in A_{i_0} \subset A$ such that $a > u$. Hence A is unbounded above.

(iii) Let $A = \bigcap_{i\in I} A_i$.

Since each A_i is bounded above and $\alpha_i = \sup A_i$, we have $a_i \leq \alpha_i$ for all $a_i \in A_i$.

If $x \in A$, then $x \in A_i$ for every $i \in I$, so $x \leq \alpha_i$ for all $i \in I$. Since every element of A is bounded above by every α_i, we conclude that if $A \neq \emptyset$, then A is bounded above by each α_i.

$A = \bigcap_{i\in I} A_i$ need not be non-empty; consider $A_n := \left(0, \frac{1}{n}\right)$ for $n \in \mathbb{N}$.

If $A = \bigcap_{i\in I} A_i \neq \emptyset$, then $\sup A \leq \alpha_i$ for all $i \in I$. Thus

$$\sup\left(\bigcap_{i\in I} A_i\right) \leq \inf_{i\in I} \alpha_i.$$

□

Exercise 14.7. *Fixed points for increasing functions.*
Let S be a non-empty ordered set such that every non-empty subset $E \subset S$ has both a supremum and an infimum. (A closed interval $[a,b]$ in $\mathbb{R}$ is an example of such an S.) Suppose $f\colon S \to S$ is monotonically increasing. Show that there exists $x \in S$ such that $f(x) = x$.

Solution. Define

$$A := \{x \in S \mid f(x) \geq x\}.$$

We consider two cases:

Case 1: $A \neq \emptyset$. Let $\alpha := \sup A$.

We claim that $f(\alpha) = \alpha$. Since f is monotonically increasing and $x \leq \alpha$ for all $x \in A$, we have

$$f(x) \leq f(\alpha) \quad \text{for all } x \in A.$$

In particular, since $f(x) \geq x$, we obtain $f(\alpha) \geq x$ for all $x \in A$; thus

$$f(\alpha) \geq \sup A = \alpha.$$

Suppose, for a contradiction, that $f(\alpha) > \alpha$. Since $f(\alpha) > \alpha$, and S is an ordered set, there exists $\alpha' \in S$ with $\alpha < \alpha' < f(\alpha)$. Since f is increasing, we have $f(\alpha') \geq f(\alpha) > \alpha'$, so $\alpha' \in A$. But this contradicts the assumption that $\alpha = \sup A$, since $\alpha' > \alpha$ and $\alpha' \in A$.

Thus, we must have $f(\alpha) \leq \alpha$. Combined with the earlier inequality $f(\alpha) \geq \alpha$, it follows that $f(\alpha) = \alpha$.

Case 2: $A = \emptyset$**.** Define

$$B := \{x \in S \mid f(x) \leq x\}.$$

Then $B \neq \emptyset$, and we can apply a symmetric argument to the infimum $\beta := \inf B$. Using similar reasoning, we conclude that $f(\beta) = \beta$.

In either case, a fixed point of f exists. □

Exercise 14.8 ([Rud76] 1.8)**.** Prove that no order can be defined in $\mathbb{C}$ that turns it into an ordered field. *Hint*: -1 is a square.

Solution. By 14.16, an order $<$ that makes $\mathbb{C}$ an ordered field would have to satisfy $-1 = i^2 > 0$, contradicting $1 > 0$. □

Exercise 14.9 ([Rud76] 1.9, lexicographic order)**.** Suppose $z = a + bi$, $w = c + di$. Define an order on $\mathbb{C}$ as follows:

$$z < w \iff \begin{cases} a < c, \text{ or} \\ a = c, b < d. \end{cases}$$

Prove that this turns $\mathbb{C}$ into an ordered set. Does this ordered set have the least upper bound property?

Solution. We show that this order turns $\mathbb{C}$ into an ordered set.

(i) Since the *real* numbers are ordered, we have $a < c$ or $a = c$ or $c < a$. In the first case $z < w$; in the third case $w < z$.

Now consider the second case where $a = c$. We must have $b < d$ or $b = d$ or $d < b$, which correspond to $z < w$, $z = w$, $w < z$ respectively.

Hence we have shown that either $z < w$ or $z = w$ or $w < z$.

(ii) We now show that if $z < w$ and $w < u$, then $z < u$. Let $u = e + fi$.

Since $z < w$, we have either $a < c$, or $a = c$ and $b < d$. Since $w < u$, we have either $c < f$, or $c = f$ and $d < g$. Hence there are four possible cases:

- $a < c$ and $c < f$. Then $a < f$ and so $z < u$, as required.
- $a < c$ and $c = f$, and $d < g$. Again $a < f$, so $z < u$.
- $a = c$, and $b < d$ and $c < f$. Once again $a < f$ so $z < u$.
- $a = c$ and $b < d$, and $c = f$ and $d < g$. Then $a = f$ and $b < g$, so $z < u$.

□

Exercise 14.10 ([Rud76] 1.10)**.** Suppose $z = a + bi$, $w = u + iv$, and

$$a = \left(\frac{|w| + u}{2}\right)^{\frac{1}{2}}, \quad b = \left(\frac{|w| - u}{2}\right)^{\frac{1}{2}}.$$

Prove that $z^2 = w$ if $v \geq 0$ and that $\overline{z}^2 = w$ if $v \leq 0$. Conclude that every complex number (with one exception!) has two complex square roots.

Solution. We have

$$a^2 - b^2 = \frac{|w| + u}{2} - \frac{|w| - u}{2} = u,$$

and

$$2ab = (|w| + u)^{\frac{1}{2}} (|w| - u)^{\frac{1}{2}} = \left(|w|^2 - u^2\right)^{\frac{1}{2}} = \left(v^2\right)^{\frac{1}{2}} = |v|.$$

Hence if $v \geq 0$,

$$z^2 = \left(a^2 - b^2\right) + 2abi = u + |v|i = w;$$

if $v \leq 0$,

$$\overline{z}^2 = \left(a^2 - b^2\right) - 2abi = u - |v|i = w.$$

Hence every non-zero w has two square roots $\pm z$ or $\pm\overline{z}$. Of course, 0 has only one square root, itself. □

Exercise 14.11 ([Rud76] 1.11)**.** If $z \in \mathbb{C}$, prove that there exists $r \geq 0$ and $w \in \mathbb{C}$ with $|w| = 1$ such that $z = rw$. Are w and r always uniquely determined by z?

Solution. If $z = 0$, take $r = 0$ and $w = 1$; in this case w is not unique.

Otherwise take $r = |z|$ and $w = \frac{z}{|z|}$; these choices are unique, since if $z = rw$, we must have $r = r|w| = |rw| = |z|$ so $w = \frac{z}{r} = \frac{z}{|z|}$ are unique. □

15 Basic Topology

Term	Notation
metric space	X, Y
points	p, q, r
metric	d
general set	E
open ball	$B_r(p)$
closed ball	$\overline{B}_r(p)$
punctured ball	$B'_r(p)$
interior	E°
closure	$\overline{E}$
boundary	∂E
induced set	E'

Table 15.1: Notation for structures in metric spaces

15.1 Metric Spaces

Definitions and Examples

Definition 15.1 (Metric space). Let X be a set. A *metric* is a function $d\colon X \times X \to [0,\infty)$ if, for all $p,q,r \in X$,

(i) $d(p,q) \geq 0$, where equality holds if and only if $p = q$; (positive definiteness)

(ii) $d(p,q) = d(q,p)$; (symmetry)

(iii) $d(p,q) \leq d(p,r) + d(r,q)$. (triangle inequality)

A ***metric space*** (X,d) is a set X together with a metric d.

For the rest of the chapter, X is taken to be a metric space, unless specified otherwise.

Example (Metrics on $\mathbb{R}^n$). Each of the following functions define metrics on $\mathbb{R}^n$.

$$d_1(x,y) = \sum_{i=1}^{n} |x_i - y_i|;$$

$$d_2(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)}$$

$$d_\infty(x,y) = \max_{i\in\{1,2,\ldots,n\}} |x_i - y_i|.$$

These are called the ℓ^1-, ℓ^2- (or Euclidean) and ℓ^∞-distances respectively.
The proof that each of d_1, d_2, d_∞ is a metric is mostly very routine.

Balls and Boundedness

Definition 15.2 (Balls).

(i) The ***open ball*** centred at $p \in X$ with radius $r > 0$ is

$$B_r(p) := \{q \in X \mid d(p,q) < r\}.$$

(ii) The ***closed ball*** centred at p with radius r is

$$\overline{B}_r(p) := \{q \in X \mid d(p,q) \leq r\}.$$

(iii) The ***punctured ball*** is the open ball excluding its centre:

$$B'_r(p) := \{q \in X \mid 0 < d(p,q) < r\}.$$

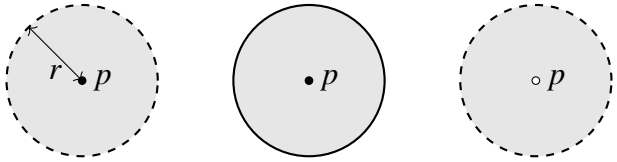


Figure 15.1: Open ball, closed ball, punctured ball

Example. Considering $\mathbb{R}^3$ with the Euclidean metric, $B_1(0)$ really is what we understand geometrically as a ball (minus its boundary, the unit sphere), whilst $\overline{B}_1(0)$ contains the unit sphere and everything inside it.

Remark. We caution that this intuitive picture of the closed ball being the open ball "together with its boundary" is totally misleading in general. For instance, in the discrete metric on a set X, the open ball $B_1(x)$ contains only the point x, whereas the closed ball $\overline{B}_1(x)$ is the whole of X.

Definition 15.3 (Bounded)**.** We say $E \subset X$ is ***bounded*** if

$$\exists M \in \mathbb{R}^+, p \in X, \quad E \subset B_M(p).$$

That is, E is contained in some open ball:

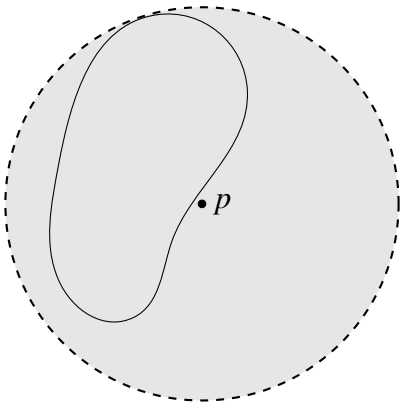


Figure 15.2: Bounded set

Proposition 15.4. *Let $E \subset X$. Then the following are equivalent:*

(i) E is bounded;

(ii) E is contained in some closed ball;

(iii) The set $\{d(p,q) \mid p,q \in E\}$ is a bounded subset of $\mathbb{R}$.

Proof.

(i) $\implies$ (ii) Suppose E is bounded. Then there exists M and $p \in X$ such that $E \subset B_M(p) \subset \overline{B}_M(p)$, which is a closed ball.

(ii) $\implies$ (iii) Suppose $E \subset \overline{B}_M(x)$ for some M and $x \in X$. Let $p,q \in E$. Then $p,q \in \overline{B}_M(x)$, so $d(p,x) \leq M$ and $d(q,x) \leq M$. By the triangle inequality,

$$d(p,q) \leq d(p,x) + d(q,x) = 2M.$$

Hence $\{d(p,q) \mid p,q \in E\}$ is bounded above by $2M$, and bounded below by 0; thus it is bounded.

(iii) $\implies$ (i) Suppose $\{d(p,q) \mid p,q \in E\}$ is a bounded subset of $\mathbb{R}$. Then there exists $M \in \mathbb{R}^+$ such that $d(p,q) \leq M$ for all $p,q \in E$. If $E = \emptyset$, then E is certainly bounded. Otherwise, let $x \in E$ be an arbitrary point. Then $E \subset B_{M+1}(x)$. □

Open and Closed Sets

Definition 15.5 (Open set)**.** We say $E \subset X$ is ***open*** (in X) if for all $p \in X$, $B_r(p) \subset E$ for some $r > 0$.

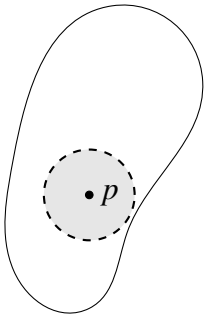
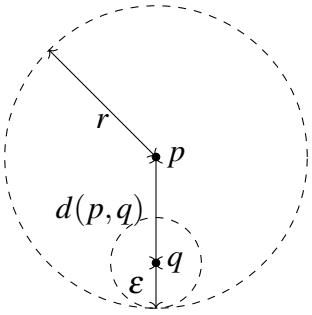


Figure 15.3: Open set

Lemma 15.6. *Any open ball is open.*



Proof. Let $B_r(p)$ be an open ball.

Let $q \in B_r(p)$. We need to show that there exists $\varepsilon > 0$ such that $B_\varepsilon(q) \subset B_r(p)$.

Take $\varepsilon := r - d(p,q)$. Let $s \in B_\varepsilon(q)$. By the triangle inequality,

$$\begin{aligned} d(p,s) &\leq d(q,s) + d(p,q) \\ &< \varepsilon + d(p,q) = r. \end{aligned}$$

Thus $s \in B_r(p)$, which implies $B_\varepsilon(q) \subset B_r(p)$. □

Lemma 15.7.

(i) Both $\emptyset$ and X are open.

(ii) For any indexing set I and collection of open sets $\{E_i\}_{i\in I}$, $\bigcup_{i\in I} E_i$ is open.

(iii) For any finite indexing set I and collection of open sets $\{E_i\}_{i\in I}$, $\bigcap_{i\in I} E_i$ is open.

Proof.

(i) Obvious by definition.

(ii) If $x \in \bigcup_{i\in I} E_i$, then $x \in E_i$ for some $i \in I$. Since E_i is open, there exists $\varepsilon > 0$ such that $B_\varepsilon(x) \subset E_i$ and hence $B_\varepsilon(x) \in \bigcup_{i\in I} E_i$.

(iii) Suppose I is finite and $x \in \bigcap_{i\in I} E_i$. For each $i \in I$, we have $x \in E_i$ and so there exists ε_i such that $B_{\varepsilon_i}(x) \subset E_i$. Set $\varepsilon = \min_{i\in I} \varepsilon_i$, then $\varepsilon > 0$ (here it is, of course, crucial that I be finite), and $B_\varepsilon(x) \subset B_{\varepsilon_i}(x) \subset E_i$ for all i. Therefore $B_\varepsilon(x) \subset \bigcap_{i\in I} E_i$.

□

While the indexing set I in (ii) can be arbitrary, the indexing set I in (iii) must be finite. **Counterexample:** $E_n = \left(-\frac{1}{n}, \frac{1}{n}\right)$ are open in $\mathbb{R}$, but their intersection $\bigcap_{n=1}^{\infty} E_n = \{0\}$ is not open.

A metric space (X,d) naturally induces a metric on any of its subsets.

Definition 15.8 (Subspace). Suppose (X,d) is a metric space, $Y \subset X$. Then the restriction of d to $Y \times Y$ gives Y a metric

$$d_Y = d|_{Y\times Y},$$

so that (Y, d_Y) is a metric space. We call Y equipped with this metric a ***subspace*** of X.

Suppose Y is a subspace of X. Let $E \subset X$ be open in X. Then E need not be open in Y.

Definition 15.9. We say E is *open relative* to Y if for all $p \in E$, $B_r(p) \cap Y \subset E$ for some $r > 0$.

Note that $B_\varepsilon(p) \cap Y$ is in the open ball in Y, because the metric $d' : Y \times Y \to [0,\infty)$ is the restriction to $Y \times Y$ of the metric $d : X \times X \to [0,\infty)$ on X.

Insert figure

The next result characterises subsets that are open relative to a subspace.

Lemma 15.10. *Suppose Y is a subspace of X, $E \subset Y$. Then E is open relative to Y if and only if $E = Y \cap G$ for some G open in X.*

Proof.

$\boxed{\Longrightarrow}$ We prove by construction.

Suppose E is open relative to Y. For each $p \in E$, there exists $r_p > 0$ such that $B_{r_p}(p) \cap Y \subset E$. Consider the union

$$\bigcup_{p \in E} \left(B_{r_p}(p) \cap Y \right) \subset E.$$

Note that

$$\bigcup_{p \in E} \left(B_{r_p}(p) \cap Y \right) = \left(\bigcup_{p \in E} B_{r_p}(p) \right) \cap Y \subset E.$$

Take

$$G := \bigcup_{p \in E} B_{r_p}(p).$$

Then we have $G \cap Y \subset E$. Since G is an intersection of open balls (which are open sets), by 15.7, G is open in X.

We now show the opposite inclusion. For each $p \in E \subset Y$, we have $p \in Y$, and $p \in B_{r_p}(p)$ for some $r_p > 0$, so $p \in \bigcup_{p \in E} B_{r_p}(p) = G$. Hence $p \in G \cap Y$. This shows $E \subset G \cap Y$.

Hence $E = G \cap Y$.

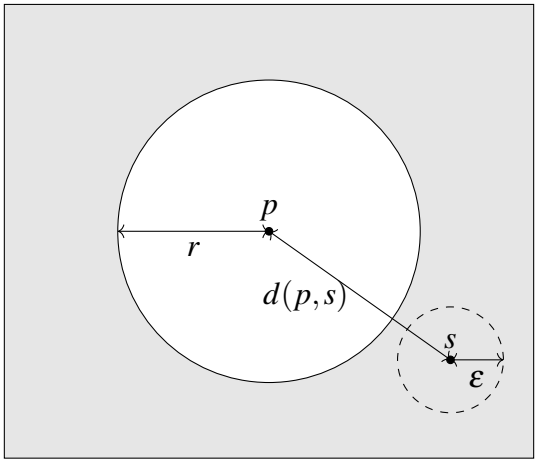
$\boxed{\Longleftarrow}$ Suppose $E = G \cap Y$ for some open subset G of X.

Let $p \in E$. Then $p \in G$. Since G is open, there exists $r_p > 0$ such that $B_{r_p}(p) \subset G$. Thus $B_{r_p}(p) \cap Y \subset G \cap Y = E$. By definition, E is open relative to Y. □

As you would have expected, the complement of an open set is a *closed* set.

Definition 15.11 (Closed set). We say $E \subset X$ is ***closed*** if its complement E^c is open.

Lemma 15.12. *Any closed ball is closed.*



Proof. To prove that $\overline{B}_r(p)$ is closed, we need to show that its complement

$$\overline{B}_r(p)^c = \{q \in X \mid d(p,q) > r\}$$

is open.

Let $s \in \overline{B}_r(p)^c$. We need to show that $B_\varepsilon(s) \subset \overline{B}_r(p)^c$ for some $\varepsilon > 0$.

Choose any $\varepsilon > 0$ such that $\varepsilon < d(p,s) - r$.

Let $q \in B_\varepsilon(s)$. Then $d(q,s) < \varepsilon < d(p,s) - r$; rearranging gives $d(p,s) - d(q,s) > r$. By the triangle inequality,

$$d(p,s) \leq d(p,q) + d(q,s),$$

so

$$d(p,q) \geq d(p,s) - d(q,s) > r.$$

Thus $q \in \overline{B}_r(p)^c$, which implies $B_\varepsilon(s) \subset \overline{B}_r(p)^c$. Therefore $\overline{B}_r(p)^c$ is open, so $\overline{B}_r(p)$ is closed. □

Lemma 15.13.

(i) Both $\emptyset$ and X are closed.

(ii) For any indexing set I and collection of closed sets $\{F_i\}_{i\in I}$, $\bigcap_{i\in I} F_i$ is closed.

(iii) For any finite indexing set I and collection of closed sets $\{F_i\}_{i\in I}$, $\bigcup_{i\in I} F_i$ is closed.

Proof. From 15.7, simply take complements and apply de Morgan's laws. □

The indexing set I in (iii) must be finite. **Counterexample:** $F_n = \left[-1 + \frac{1}{n}, 1 - \frac{1}{n}\right]$ are closed in $\mathbb{R}$, but their union $\bigcup_{n=1}^{\infty} F_n = (-1,1)$ is open.

Interior, Closure, Boundary

Definition 15.14. Suppose $E \subset X$.

(i) The ***interior*** E° of the set E is the union of all open subsets of X contained in E; we call $p \in E^\circ$ an *interior point* of E.

$$E^\circ := \bigcup_{\substack{F \subset E, \\ F \text{ is open}}} F.$$

(ii) The ***closure*** $\overline{E}$ of the set E is the intersection of all closed subsets of X containing E.

$$\overline{E} := \bigcap_{\substack{F \supset E, \\ F \text{ is closed}}} F.$$

We say E is ***dense*** if $\overline{E} = X$.

(iii) The ***boundary*** of E is $\partial E := \overline{E} \setminus E^\circ$; we call $p \in \partial E$ a ***boundary point*** of E.

From the definitions, the interior is the *largest* open set contained in E; the closure is the *smallest* closed set containing E.

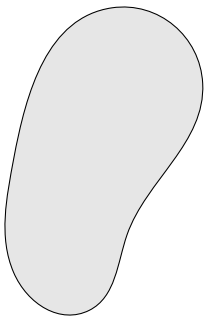


Figure 15.4: Interior, closure, boundary

Example.

- The interior of the closed interval $[a,b]$ is the open interval (a,b).
- $\mathbb{Q}$ is dense in $\mathbb{R}$.

Remark. E and E° do not necessarily have the same closures. **Counterexample:** take $E = \mathbb{Q}$, then $\overline{E} = \mathbb{R}$ and $\overline{E^\circ} = \emptyset$.

Likewise, E and $\overline{E}$ do not necessarily have the same interiors. **Counterexample:** take $E = (-1,0) \cup (0,1) \subset \mathbb{R}$. Then $E^\circ = (-1,0) \cup (0,1)$ and $(\overline{E})^\circ = [-1,1]$.

Lemma 15.15. *Suppose $E \subset X$.*

(i) E is open $\iff E = E^\circ$.

(ii) E is closed $\iff E = \overline{E}$.

Proof.

(i) $\boxed{\Longrightarrow}$ Suppose E is open. Since $E \subset E$ and E is open by assumption, $E \subset E^\circ$.

To show the opposite containment, let $x \in E^\circ$. Then $x \in F$ for some $F \subset E$, F is open. Thus $x \in E$. Hence $E^\circ \subset E$.

$\boxed{\Longleftarrow}$ Since an arbitrary union of open sets is open, E° is open. Since $E = E^\circ$, E is open.

(ii) $\boxed{\Longrightarrow}$ Suppose E is closed. Evidently $E \subset \overline{E}$.

To show the opposite containment, let $x \in \overline{E}$. Then x is in every closed subset of X containing E, so $x \in E$. Hence $\overline{E} \subset E$.

$\boxed{\Longleftarrow}$ Since an arbitrary intersection of closed sets is closed, $\overline{E}$ is closed. Since $E = \overline{E}$, E is closed.

□

Proposition 15.16. *Suppose $E \subset X$. Then $p \in \overline{E}$ if and only if every open ball centred at p contains a point of E.*

Proof.

$\boxed{\Longrightarrow}$ Let $p \in \overline{E}$.

Suppose, for a contradiction, that there exists an open ball $B_\varepsilon(p)$ that does not meet E. Then $B_\varepsilon(p)^c$ is a closed set containing E. Therefore $B_\varepsilon(p)^c$ contains $\overline{E}$, and hence it contains p, which is obviously nonsense.

$\boxed{\Longleftarrow}$ Suppose for each $\varepsilon > 0$, $B_\varepsilon(p) \cap E \neq \emptyset$.

Suppose, for a contradiction, that $p \notin \overline{E}$. Since $\overline{E}^c$ is open, there is a ball $B_\varepsilon(p)$ contained in $\overline{E}^c$, and hence in E^c, contrary to assumption. □

Corollary 15.17. *Suppose $E \subset X$. Then E is dense if and only if it meets every open set in X.*

The next result summarises basic properties of closure and interior.

Lemma 15.18. *Suppose $A, B \subset X$. Then*

(i) $\overline{A \cup B} = \overline{A} \cup \overline{B}$

(ii) $\overline{A \cap B} \subset \overline{A} \cap \overline{B}$

(iii) $(A \cup B)^\circ \supset A^\circ \cup B^\circ$

(iv) $(A \cap B)^\circ = A^\circ \cap B^\circ$

(v) $(A^\circ)^c = \overline{A^c}$

(vi) $(\overline{A})^c = (A^c)^\circ$

Proof.

(i)

□

Limit Points

Definition 15.19 (Limit point)**.** We say $p \in X$ is a ***limit point*** of E if

$$\forall \varepsilon > 0, \quad \exists q \in E \setminus \{p\}, \quad q \in B_\varepsilon(p).$$

That is, every open ball centred at p has a non-empty intersection with the set E:

$$(B_\varepsilon(p) \setminus \{p\}) \cap E \neq \emptyset \quad \text{for all } \varepsilon > 0.$$

Remark. We do not require a limit point of a set to belong to the set.

The *induced set* of E, denoted by E', is the set of all limit points of E in X.

We say $p \in E$ is an *isolated point* of E if p is not an limit point of E (that is, there exists $\varepsilon > 0$ such that $B_\varepsilon(p) \cap E = \{p\}$).

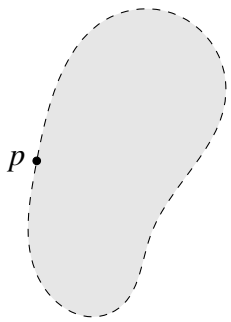


Figure 15.5: Limit point

Example.

- The set $\left\{\frac{1}{n} \,\middle|\, n \in \mathbb{N}\right\}$ has 0 as a limit point.
- The set of rational numbers has every real number as a limit point.
- Every point of $[a,b]$ is a limit point of the set of numbers in (a,b).
- Consider $\mathbb{R}^2$. The set of limit points of any open ball $B_r(p)$ is the closed ball $\overline{B}_r(p)$, which is also the closure of $B_r(p)$.
- Consider $\mathbb{Q} \subset \mathbb{R}$. $\mathbb{Q}' = \overline{\mathbb{Q}} = \mathbb{R}$.

Proposition 15.20. *If p is a limit point of E, then every open ball of p contains infinitely many points of E.*

Proof. Suppose, for a contradiction, that there exists an open ball $B_r(p)$ which contains only a finite number of points of E distinct from p; let

$$B_r(p) = \{q_1, \ldots, q_n\},$$

where $p \neq q_i$ for $i = 1, \ldots, n$. Choose

$$r := \min_{1 \leq i \leq n} d(p, q_i).$$

Then $B_r(p)$ contains no points of E distinct from p, which is a contradiction. □

Corollary 15.21. *A finite point set has no limit points.*

The converse is not true. **Counterexample:** $\mathbb{N}$ is an infinite set with no limit points.

In a later section, we will show that infinite sets contained in some open ball always have an limit point; this result is known as the Bolzano–Weierstrass theorem (15.41).

A closed set was defined to be the complement of an open set. The next result characterises closed sets in another way.

Lemma 15.22. *Suppose $E \subset X$. Then E is closed if and only if it contains all its limit points.*

Proof.

$\boxed{\Longrightarrow}$ Suppose E is closed. Let p be a limit point of E. We want to show $p \in E$.

Suppose, for a contradiction, that $p \notin E$. Then $p \in E^c$. Since E^c is open, there exists $\varepsilon > 0$ such that $B_\varepsilon(p) \subset E^c$. Thus $B_\varepsilon(p)$ does not intersect E, so p is not a limit point of E.

$\boxed{\Longleftarrow}$ Suppose E contains all its limit points. To show that E is closed, we need to show that E^c is open.

Let $p \in E^c$. Then p is not a limit point of E, so there exists some ball $B_\varepsilon(p)$ which does not intersect E, so $B_\varepsilon(p) \subset E^c$. Hence E^c is open, so E is closed. □

Lemma 15.23. *Suppose $E \subset X$. Then E' is a closed subset of X.*

Proof. To prove that E' is closed, we will show its complement $(E')^c$ is open.

Let $p \in (E')^c$. Then $p \notin E'$, so p is not a limit point of E; thus, there exists a ball $B_\varepsilon(p)$ whose intersection with E is either empty or $\{p\}$ (depending on whether $p \in E$ or not).

We will show that $B_{\frac{\varepsilon}{2}}(p) \subset (E')^c$. Let $q \in B_{\frac{\varepsilon}{2}}(p)$.

Case 1: $q = p$**.** Then clearly $q \in (E')^c$.

Case 2: $q \neq p$**.** There is some ball about q which is contained in $B_\varepsilon(p)$, but does not contain p: the ball $B_\delta(q)$ where $\delta = \min\left(\frac{\varepsilon}{2}, d(p,q)\right)$ has this property. This ball meets E in the empty set, and so $q \in (E')^c$ in this case too.

□

The next result provides a useful expression for the closure of a set; it states that every point of $\overline{E}$ is either a limit point of E, or in E.

Lemma 15.24. *Suppose $E \subset X$. Then $\overline{E} = E \cup E'$.*

Proof. We show double inclusion.

$\boxed{\supset}$ Since $E \subset \overline{E}$, it suffices to show that $E' \subset \overline{E}$.

We prove the contrapositive. Let $p \notin \overline{E}$. Then $p \in \overline{E}^c$. Since $\overline{E}^c$ is open, there exists $\varepsilon > 0$ such that $B_\varepsilon(p) \subset \overline{E}^c$, so $B_\varepsilon(p) \subset E^c$. This implies $B_\varepsilon(p) \cap E = \emptyset$, so p is not a limit point of E. Thus $p \notin E'$.

$\boxed{\subset}$ If $p \in \overline{E}$, we saw in Lemma 5.1.5 that there is a sequence (x_n) of elements of E with $x_n \to p$. If $x_n = p$ for some n then we are done, since this implies that $p \in E$. Suppose, then, that $x_n \neq p$ for all n. Let $\varepsilon > 0$ be given, for sufficiently large n, all the x_n are elements of $B_\varepsilon(p) \setminus \{p\}$, and they all lie in E. It follows that p is a limit point of E, and so we are done in this case also. □ to do

Lemma 15.25. *Suppose non-empty $E \subset \mathbb{R}$ is bounded above. Then* $\sup E \in \overline{E}$.
If E is closed, then $\sup E \in E$.

Proof. Let $y = \sup E$.

Case 1: $y \in E$**.** Since $E \subset \overline{E}$, we have $y \in \overline{E}$.

Case 2: $y \notin E$**.** For every $\varepsilon > 0$, there exists $x \in E$ such that $y - \varepsilon < x < y$ (for otherwise $y - \varepsilon$ would be an upper bound of E). Thus every open interval $(y - \varepsilon, y + \varepsilon)$ centred at y intersects E, so y is a limit point of E. Hence $y \in \overline{E}$.

□

15.2 Compactness

Definitions and Properties

Definition 15.26 (Open cover)**.** An ***open cover*** of $K \subset X$ is a collection of open sets $\mathscr{U} = \{U_i\}_{i \in I}$ such that

$$K \subset \bigcup_{i \in I} U_i.$$

We say $\mathscr{V} \subset \mathscr{U}$ is a *subcover* of $\mathscr{U}$; if $\mathscr{V}$ is finite, we call $\mathscr{V}$ a *finite* subcover.

Definition 15.27 (Compactness)**.** We say $K \subset X$ is ***compact*** if *every* open cover of K contains a finite subcover.

That is, if $\mathscr{U} = \{U_i\}_{i \in I}$ is an open cover of K, there exists finitely many indices $i_1, \ldots, i_n \in I$ such that

$$K \subset \bigcup_{k=1}^{n} U_{i_k}.$$

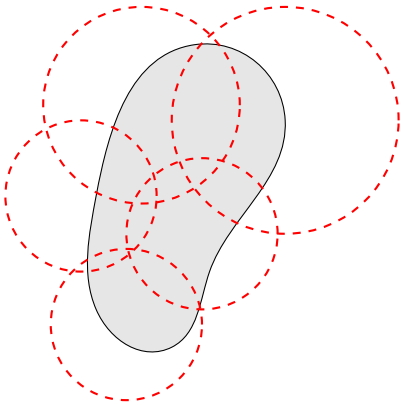


Figure 15.6: Compact set

Example.

- $\mathbb{R}$ is not compact, since the open cover $\{(-n,n) \mid n \in \mathbb{N}\}$ has no finite subcover.
- $\mathbb{Z}$ is not compact in $\mathbb{R}$, since the open cover $\left\{\left(n-\frac{1}{2}, n+\frac{1}{2}\right) \,\middle|\, n \in \mathbb{Z}\right\}$ has no finite subcover.
- $[0,1]$ is compact. (See 15.32 for the proof.)

Lemma 15.28. *Every finite set is compact.*

Proof. Let $E = \{p_1, \ldots, p_n\}$. Let $\mathscr{U} = \{U_i\}_{i \in I}$ be an open cover of E.

For each $p_k \in E$, choose *one* U_{i_k} such that $p_k \in U_{i_k}$. Then $\{U_{i_1}, \ldots, U_{i_n}\}$ is a finite subcover of $\mathscr{U}$. □

Notice earlier than if $E \subset Y \subset X$, then E may be open relative to Y, but not open relative to X; this implies that the property of being open depends on the space in which E is embedded. Compactness, however, behaves better, as shown in the next result; it is independent of the metric space.

Proposition 15.29. *Suppose Y is a subspace of X, and $K \subset Y$. Then K is compact relative to X if and only if K is compact relative to Y.*

Proof.

$\boxed{\Longrightarrow}$ Suppose K is compact relative to X.

Let $\mathscr{U} = \{U_i\}_{i \in I}$ be an open cover of K in Y.

Since each U_i is open relative to Y, by 15.10, there exists V_i open relative to X such that $U_i = Y \cap V_i$. Consider the open cover $\{V_i\}_{i \in I}$ of K. Since K is compact relative to X, there exist finitely many indices $i_1, \ldots, i_n$ such that

$$K \subset \bigcup_{k=1}^{n} V_{i_k}.$$

Since $K \subset \bigcup_{k=1}^{n} V_{i_k}$ and $K \subset Y$, we have that

$$K \subset \left(\bigcup_{k=1}^{n} V_{i_k} \right) \cap Y = \bigcup_{k=1}^{n} (Y \cap V_{i_k}) = \bigcup_{k=1}^{n} U_{i_k},$$

where $\{U_{i_1}, \ldots, U_{i_n}\}$ forms a finite subcover of $\mathscr{U}$. Hence K is compact relative to Y.

$\boxed{\Longleftarrow}$ Suppose K is compact relative to Y. Let $\mathscr{V} = \{V_i\}_{i \in I}$ be an open cover of K in X.

For $i \in I$, let $U_i = Y \cap V_i$. Then $\{U_i\}_{i \in I}$ cover K in Y. By compactness of K in Y, there exist finitely many indices $i_1, \ldots, i_n$ such that

$$K \subset \bigcup_{k=1}^{n} U_{i_k} \subset \bigcup_{k=1}^{n} V_{i_k}$$

since $U_i \subset V_i$. Hence $\{V_{i_1}, \ldots, V_{i_n}\}$ is a finite subcover of $\mathscr{V}$, so K is compact relative to X. □

Proposition 15.30. *Compact subsets of metric spaces are bounded.*

Proof. Suppose $K \subset X$ is compact. To prove that K is bounded, we want to construct some open ball that contains the entirety of K.

Fix $p \in K$. For $n \in \mathbb{N}$, let $U_n = B_n(p)$. Then $\{U_n\}_{n\in\mathbb{N}}$ is an open cover of K. By compactness of K, there exists a finite subcover

$$\{U_{n_1}, \ldots, U_{n_m}\}.$$

But $U_{n_1} \subset \cdots \subset U_{n_m}$, so U_{n_m} contains K. Hence K is bounded. □

Proposition 15.31. *Compact subsets of metric spaces are closed.*

Proof. Let $K \subset X$ be compact. To prove that K is closed, we need to show that K^c is open. Let $p \in K^c$; our goal is to show that there exists $\varepsilon > 0$ such that $B_\varepsilon(p) \subset K^c$, or $B_\varepsilon(p) \cap K = \emptyset$.

For all $q_i \in K$, consider the pair of open balls $B_{r_i}(p)$ and $B_{r_i}(q_i)$, where $r_i < \frac{1}{2}d(p,q_i)$. Since K is compact, there exists finite many points $q_{i_1}, \ldots, q_{i_n} \in K$ such that

$$K \subset \bigcup_{k=1}^{n} B_{r_{i_k}}(q_{i_k}) = W.$$

Consider the intersection

$$\bigcap_{k=1}^{n} B_{r_{i_k}}(p),$$

which is an open ball at p of radius $\min\limits_{1\le k\le n} d(p,q_{i_k})$.

Pick $\varepsilon := \min\limits_{1\le k\le n} d(p,q_{i_k})$.

Note that $B_\varepsilon(p) \subset B_{r_{i_k}}(p)$ for all $k = 1, \ldots, n$. By construction, for all $q_i \in K$, the open balls $B_{r_i}(p)$ and $B_{r_i}(q_i)$ are disjoint. In particular,

$$B_\varepsilon(p) \cap B_{r_{i_k}}(q_{i_k}) = \emptyset \quad (k = 1, \ldots, n)$$

Then

$$B_\varepsilon(p) \cap W = B_\varepsilon(p) \cap \left(\bigcup_{k=1}^{n} B_{r_{i_k}}(q_{i_k}) \right) = \bigcup_{k=1}^{n} \left(B_\varepsilon(p) \cap B_{r_{i_k}}(q_{i_k}) \right) = \emptyset$$

as desired. □

Proposition 15.32. *Closed subsets of compact sets are compact.*

Proof. Suppose $K \subset X$ is compact, $F \subset K$ is closed (relative to X). We will show that F is compact. Let $\mathscr{U} = \{U_i\}_{i\in I}$ be an open cover of F. We will construct a finite subcover of $\mathscr{U}$.

Since F is closed, its complement F^c is open. Consider the union

$$\Omega = \mathscr{U} \cup \{F^c\},$$

which is an open cover of K.

Since K is compact, there exists a finite subcover of Ω, given by

$$\Phi = \{U_{i_1}, \ldots, U_{i_n}, F^c\}$$

which covers K, and hence F. Now remove F^c from Φ to obtain

$$\Phi' = \{U_{i_1}, \ldots, U_{i_n}\},$$

which is an open cover of F, since $F^c \cap F = \emptyset$. Hence Φ' is a finite subcover of $\mathscr{U}$, so F is compact. □

Remark. This does *not* say "closed sets are compact"! In fact, closed sets are not necessarily compact. **Counterexample:** $\mathbb{R}$ is closed in $\mathbb{R}$, but it is not compact because it is not bounded.

Note that closed and bounded sets are not necessarily compact for general metric spaces, but they are compact in $\mathbb{R}^n$ (see 15.40).

Corollary 15.33. *If F is closed and K is compact, then $F \cap K$ is compact.*

Proof. Suppose F is closed, K is compact. By 15.31, K is closed. By 15.13, the intersection of two closed sets is closed, so $F \cap K$ is closed.

Since $F \cap K \subset K$ is a closed subset of a compact set K, by 15.32, $F \cap K$ is compact. □

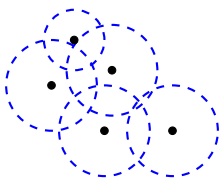
Heine–Borel Theorem

Proposition 15.34. *K is compact if and only if every infinite subset of K has a limit point in K.*

Proof.

$\boxed{\Longrightarrow}$ Suppose K is compact. Let E be an infinite subset of K. Suppose, for a contradiction, that E has no limit point in K.

For all $p \in K$, p is not a limit point of E, so there exists $r_p > 0$ such that $B_{r_p}(p) \cap E \setminus \{p\} = \emptyset$.



Consider the open cover of K given by the collection of open balls at each $p \in K$:

$$\mathscr{U} = \left\{B_{r_p}(p) \mid p \in E\right\}.$$

It is clear that $\mathscr{U}$ has no finite subcover, since E is infinite, and each $B_{r_p}(p)$ contains at most one point of E.

Since $E \subset K$, the above is also true for K. This contradicts the compactness of K.

$\Longleftarrow$ Suppose every infinite subset of K that has a limit point in K. Fix an arbitrary open cover $\mathscr{U} = \{U_i\}_{i \in I}$ of K. We will show that $\mathscr{U}$ has a finite subcover, by construction.

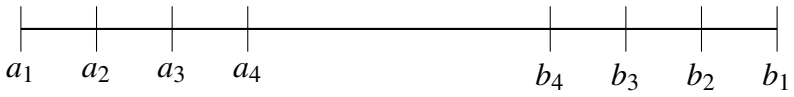
Before that, we will reindex $\mathscr{U}$ to make it more convenient, as follows. By the definition of a cover, every $p \in K$ is contained in some U_i. Pick *one* such U_i for each $p \in K$, and call it U_p. Then our open cover is now $\mathscr{U} = \{U_p \mid p \in K\}$, and for all $p \in K$ we have $p \in U_p$.

To complete proof

□

Proposition 15.35 (Nested interval theorem)**.** *Suppose (I_n) is a decreasing sequence of closed and bounded intervals in $\mathbb{R}$; that is, $I_1 \supset I_2 \supset \cdots$. Then*

$$\bigcap_{n=1}^{\infty} I_n \neq \emptyset.$$



Proof. Let $I_n = [a_n, b_n]$, for $n = 1, 2, \ldots$.

Let $E = \{a_n \mid n \in \mathbb{N}\}$. Since E is non-empty and bounded above (by b_1), it has a supremum in $\mathbb{R}$; let $x = \sup E$.

Claim. $x \in \bigcap_{n=1}^{\infty} I_n$.

Since x is the supremum, $a_n \leq x$ for all $n \in \mathbb{N}$. Note that for $m > n$, $I_n \supset I_m$ implies $a_n \leq a_m \leq b_m \leq b_n$. This means b_n is an upper bound for all a_n; hence $x \leq b_n$ for all $n \in \mathbb{N}$.

Therefore $x \in I_n$ for $n = 1, 2, \ldots$. □

To generalise the notion of intervals, we define a *k-cell* as

$$\{(x_1, \ldots, x_k) \in \mathbb{R}^k \mid a_i \leq x_i \leq b_i,\ 1 \leq i \leq k\}.$$

Example. A 1-cell is an interval, a 2-cell is a rectangle, and a 3-cell is a rectangular solid. In this regard, we can think of a k-cell as a higher-dimensional version of a rectangle or rectangular solid; it is the Cartesian product of k closed intervals.

The previous result can be generalised to k-cells, which we will now prove.

Proposition 15.36. *Suppose (I_n) is a decreasing sequence of k-cells; that is, $I_1 \supset I_2 \supset \cdots$. Then $\bigcap_{n=1}^{\infty} I_n \neq \emptyset$.*

Proof. Let I_n consist of all points $\mathbf{x} = (x_1, \ldots, x_k)$ such that

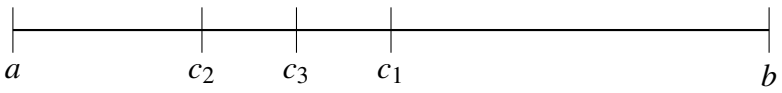
$$a_{n,i} \leq x_i \leq b_{n,i} \quad (1 \leq i \leq k; n = 1, 2, \ldots),$$

and put $I_{n,i} = [a_{n,i}, b_{n,i}]$. For each i, the sequence $(I_{n,i})$ satisfies the hypotheses of 15.35. Hence there are real numbers x_i' $(1 \leq i \leq k)$ such that

$$a_{n,i} \leq x_i' \leq b_{n,i} \quad (1 \leq i \leq k; n = 1, 2, \ldots).$$

Setting $\mathbf{x}' = (x_1', \ldots, x_k')$, we see that $\mathbf{x}' \in I_n$ for $n = 1, 2, \ldots$. Hence $\bigcap_{n=1}^{\infty} I_n \neq \emptyset$, as desired. □

Lemma 15.37. *Every closed interval is compact (in $\mathbb{R}$).*



Proof. Suppose, for a contradiction, that a closed interval $[a, b] \subset \mathbb{R}$ is not compact. Then there exists an open cover $\mathscr{U} = \{U_i\}_{i \in I}$ with no finite subcover.

Let $c_1 = \frac{1}{2}(a, b)$. Subdivide $[a, b]$ into subintervals $[a, c_1]$ and $[c_1, b]$. Then $\mathscr{U}$ covers $[a, c_1]$ and $[c_1, b]$, but at least one of these subintervals has no finite subcover (if not, then both subintervals have finite subcovers, so we can take the union of the two finite subcovers to obtain a larger subcover of the entire interval). WLOG, assume $[a, c_1]$ has no finite subcover; let $I_1 = [a, c_1]$.

Again subdivide I_1 in half to get $[a, c_2]$ and $[c_2, c_1]$. At least one of these subintervals has no finite subcover.

Repeat the above process of subdividing intervals into half. Then we obtain a decreasing sequence of closed intervals

$$I_1 \supset I_2 \supset I_3 \supset \cdots$$

where all of them have no finite subcover of $\mathscr{U}$.

By the nested interval theorem (15.35), there exists $x' \in I_n$ for all $n \in \mathbb{N}$. Notice x' is in some U_i, which is open. Then there exists $\varepsilon > 0$ such that $B_\varepsilon(x') \subset U_i$.

Since the length of the subintervals is decreasing and tends to zero, there exists some subinterval I_n so small such that $I_n \subset B_\varepsilon(x')$. This means $I_n \subset U_i$, so U_i itself is an open cover of I_n, which contradicts the fact that I_n has no finite subcover of $\mathscr{U}$. □

We now show a more general result.

Lemma 15.38. *Every k-cell is compact (in $\mathbb{R}^k$).*

Proof. We proceed in a similar manner to the proof the previous result.

Suppose I is a k-cell; that is,

$$I = \{(x_1, \ldots, x_k) \mid a_i \leq x_i \leq b_i,\ 1 \leq i \leq k\}.$$

Write $\mathbf{x} = (x_1, \ldots, x_k) \in \mathbb{R}^k$. Let

$$\delta = \left(\sum_{i=1}^{k} (b_i - a_i)^2 \right)^{1/2}$$

that is, the distance between the points $(a_1, \ldots, a_k)$ and $(b_1, \ldots, b_k)$, which is the maximum distance between two points in I: for all $\mathbf{x}, \mathbf{y} \in I$,

$$|\mathbf{x} - \mathbf{y}| \leq \delta.$$

Suppose, for a contradiction, that I is not compact; that is, there exists an open cover $\mathscr{U} = \{U_i\}_{i \in I}$ of I which contains no finite subcover of I.

For $1 \leq i \leq k$, let $c_i = \frac{1}{2}(a_i + b_i)$. The intervals $[a_i, c_i]$ and $[c_i, b_i]$ then determine 2^k k-cells Q_i whose union is I. At least one of these sets Q_i, call it I_1, cannot be covered by any finite subcollection of $\mathscr{U}$ (otherwise I could be so covered). We next subdivide I_1 and continue the process. We obtain a sequence (I_n) with the following properties:

(i) $I \supset I_1 \supset I_2 \supset \cdots$

(ii) I_n is not covered by any finite subcollection of $\mathscr{U}$

(iii) $|\mathbf{x} - \mathbf{y}| \leq \dfrac{\delta}{2^n}$ for all $\mathbf{x}, \mathbf{y} \in I_n$

By (i) and 15.36, there is a point $\mathbf{x}'$ which lies in every I_n. For some i, $\mathbf{x}' \in U_i$. Since U_i is open, there exists $r > 0$ such that $|\mathbf{y} - \mathbf{x}'| < r$ implies that $y \in U_i$. If n is so large that $2^{-n}\delta < r$ (there is such an n, for otherwise $2^n \leq \frac{\delta}{r}$ for all positive integers n, which is absurd since $\mathbb{R}$ is archimedean), then (iii) implies that $I_n \subset U_i$, which contradicts (ii). □

We have now come to an important result, which will be crucial in proving the Heine–Borel theorem and Bolzano–Weierstrass theorem.

Proposition 15.39. *If $E \subset \mathbb{R}^k$ has one of the following three properties, then it has the other two:*

(i) E is closed and bounded.

(ii) E is compact.

(iii) Every infinite subset of E has a limit point in E.

Proof.

(i) $\Longrightarrow$ (ii) Suppose E is closed and bounded. Since E is bounded, then $E \subset I$ for some k-cell I.

By 15.38, I is compact. Since E is a closed subset of a compact set, by 15.32, E is compact.

(ii) $\Longrightarrow$ (iii) This directly follows from 15.34.

(iii) $\Longrightarrow$ (i) If E is not bounded, then E contains points $\mathbf{x}_n$ with

$$|\mathbf{x}_n| > n \quad (n = 1,2,3,\dots)$$

The set S consisting of these points $\mathbf{x}_n$ is infinite and clearly has no limit point in $\mathbb{R}^k$, hence has none in E. Thus (iii) implies that E is bounded.

If E is not closed, then there is a point $\mathbf{x}_0 \in \mathbb{R}^k$ which is a limit point of E but not a point of E. For $n = 1,2,3,\dots$, there are points $\mathbf{x}_n \in E$ such that $|\mathbf{x}_n - \mathbf{x}_0| < \frac{1}{n}$. Let S be the set of these points $\mathbf{x}_n$. Then S is infinite (otherwise $|\mathbf{x}_n - \mathbf{x}_0|$ would have a constant positive value, for infinitely many n), S has $\mathbf{x}_0$ as a limit point, and S has no other limit point in $\mathbb{R}^k$. For if $\mathbf{y} \in \mathbb{R}^k$, $\mathbf{y} \neq \mathbf{x}_0$, then

$$\begin{aligned} |\mathbf{x}_n - \mathbf{y}| &\geq |\mathbf{x}_0 - \mathbf{y}| - |\mathbf{x}_n - \mathbf{x}_0| \\ &\geq |\mathbf{x}_0 - \mathbf{y}| - \frac{1}{n} \\ &\geq \frac{1}{2}|\mathbf{x}_0 - \mathbf{y}| \end{aligned}$$

for all but finitely many n; this shows that $\mathbf{y}$ is not a limit point of S (Theorem 2.20).

Thus S has no limit point in E; hence E must be closed if (iii) holds. □ review proof

Corollary 15.40 (Heine–Borel theorem). *$E \subset \mathbb{R}^n$ is compact if and only if E is closed and bounded.*

Proof. This is simply (i) $\Longleftrightarrow$ (ii) in the previous result. □

Bolzano–Weierstrass Theorem

Theorem 15.41 (Bolzano–Weierstrass theorem). *Every bounded infinite subset of $\mathbb{R}^n$ has a limit point in $\mathbb{R}^n$.*

Proof. Suppose E is a bounded infinite subset of $\mathbb{R}^n$.

Since E is bounded, there exists an n-cell $I \subset \mathbb{R}^n$ such that $E \subset I$. Since I is compact, by 15.34, E has a limit point in I and thus $\mathbb{R}^n$. □

Cantor's Intersection Theorem

A collection $\mathscr{A}$ of subsets of X is said to have the *finite intersection property* if the intersection of every finite subcollection of $\mathscr{A}$ is non-empty.

Proposition 15.42. *Suppose $\mathscr{K} = \{K_i\}_{i\in I}$ is a collection of compact subsets of a metric space X, which satisfies the finite intersection property. Then $\bigcap_{i\in I} K_i \neq \emptyset$.*

Proof. We fix a member $K_1 \subset \mathscr{K}$. Suppose, for a contradiction, that $\bigcap_{i\in I} K_i = \emptyset$; that is, no point of K_1 belongs to every $K_i \in \mathscr{K}$.

For $i \in I$, let $U_i = K_i{}^c$. Then the sets $\{U_i\}_{i\in I}$ form an open cover of K_1. Since K_1 is compact by assumption, there exist finitely many indices $i_1, \ldots, i_n$ such that

$$K_1 \subset \bigcup_{k=1}^{n} U_{i_k}.$$

By de Morgan's laws, we have that

$$\bigcup_{k=1}^{n} U_{i_k} = \bigcup_{k=1}^{n} K_{i_k}{}^c = \left(\bigcap_{k=1}^{n} K_{i_k}\right)^c.$$

Thus

$$K_1 \subset \left(\bigcap_{k=1}^{n} K_{i_k}\right)^c,$$

which means that

$$K_1 \cap \bigcap_{k=1}^{n} K_{i_k} = \emptyset.$$

Thus $K_1, K_{i_1}, \ldots, K_{i_n}$ is a finite subcollection of $\mathscr{K}$ which has an empty intersection; this contradicts the finite intersection property of $\mathscr{K}$. □

Corollary 15.43 (Cantor's intersection theorem). *Suppose (K_n) is a decreasing sequence of non-empty compact sets; that is, $K_1 \supset K_2 \supset \cdots$. Then $\bigcap_{n=1}^{\infty} K_n \neq \emptyset$.*

Proof. This is an immediate corollary of the previous result. □

The following result is a characterisation of compact sets.

Proposition 15.44. *K is compact if and only if every collection of closed subsets of K satisfies the finite intersection property.*

Proof.

$\boxed{\Longrightarrow}$ Suppose K is compact.

If $\mathscr{U}$ is an open covering of K, then the collection $\mathscr{F}$ of complements of sets in $\mathscr{U}$ is a collection of closed sets whose intersection is empty (why?); and

conversely, if $\mathscr{F}$ is a collection of closed sets whose intersection is empty, then the collection $\mathscr{U}$ of complements of sets in $\mathscr{F}$ is an open covering.

To complete proof

□

Sequential Compactness

Definition 15.45 (Sequential compactness)**.** We say $K \subset X$ is ***sequentially compact*** if every sequence in K has a convergent subsequence in K.

We now show that compactness and sequential compactness are equivalent.

Proposition 15.46. *$K \subset X$ is compact if and only if it is sequentially compact.*

Proof.

$\boxed{\Longrightarrow}$ Suppose $K \subset X$ is compact. Take any sequence (y_n) from K. Suppose, for a contradiction, that every point $x \in K$ is not a limit of any subsequence of (y_n). Then for all $x \in K$, there exists $r_x > 0$ such that $B_{r_x}(x)$ contains at most one point in (y_n), which is x.

Consider the collection of open balls at each $x \in K$:

$$\{B_{r_x}(x) \mid x \in K\}.$$

This is an open cover of K. By the compactness of K, there exists a finite subcover of K:

$$\left\{B_{r_{x_1}}(x_1), \ldots, B_{r_{x_N}}(x_N)\right\}.$$

In particular, these open balls cover $\{y_n\}$. Hence there must be some x_i $(1 \leq i \leq N)$ such that there are infinitely many $y_j = x_i$. Consider the sequence (y_j) where each term in this sequence is equal to x_i; this is a subsequence of (y_n) that converges to $x_i \in K$. This contradicts the assumption.

$\boxed{\Longleftarrow}$ Suppose, for a contradiction, that K is not compact. Then there exists an open cover $\{U_\alpha \mid \alpha \in \Lambda_\alpha\}$ which has no finite subcover. Then Λ must be an infinite set.

If Λ is countable, WLOG, assume $\Lambda = \mathbb{N}$. Since any finite union

$$\bigcup_{i=1}^{n} U_i$$

cannot cover K, we can take some $x_n \in K \setminus \bigcup_{i=1}^{n} U_i$ for every $n \in \mathbb{N}$. Then we obtain a sequence (x_n) in K and so must have a convergent subsequence (x_{n_k}) that converges to some $x_0 \in K$. It follows that there must be some U_N such that $x_0 \in U_N$. Since U_N is open, there exists $r > 0$ such that

$$B_r(x_0) \subset U_N.$$

On the other hand, since $x_{n_k} \to x_0$, there exists $N' \in \mathbb{N}$ such that if $n_k \geq N'$ then

$$x_{n_k} \in B_r(x_0).$$

However, by our way of choosing x_n, whenever $n_k > \max\{N', N\}$, $x_{n_k} \notin U_N$. This leads to a contradiction. □

15.3 Perfect Sets

Definition and Uncountability

Definition 15.47 (Perfect set). E is ***perfect*** if

(i) E is closed, and

(ii) every point of E is a limit point of E.

Proposition 15.48. *Let non-empty $P \subset \mathbb{R}^k$ be perfect. Then P is uncountable.*

Proof. Since P has limit points, by 15.20, P is an infinite set.

Suppose, for a contradiction, that P is countable. This means we can list the points of P in a sequence:

$$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots$$

Consider a sequence (B_n) of open balls, where B_n is any open ball centred at $\mathbf{x}_n$:

$$B_n = \left\{ \mathbf{y} \in \mathbb{R}^k \,\middle|\, |\mathbf{y} - \mathbf{x}_n| < r \right\}.$$

Then its closure $\overline{B_n}$ is the closed ball

$$\overline{B}_n = \left\{ \mathbf{y} \in \mathbb{R}^k \,\middle|\, |\mathbf{y} - \mathbf{x}_n| \leq r \right\}.$$

Suppose B_n has been constructed. Note that $B_n \cap P$ is not empty. Since P is perfect, every point of P is a limit point of P, so there exists B_{n+1} such that (i) $\overline{B}_{n+1} \subset B_n$, (ii) $\mathbf{x}_n \notin \overline{B}_{n+1}$, (iii) $B_{n+1} \cap P$ is not empty.

By (iii), B_{n+1} satisfies our induction hypothesis, and the construction can proceed.

Put $K_n = \overline{B}_n \cap P$. Since $\overline{B}_n$ is closed and bounded, $\overline{B}_n$ is compact. Since $\mathbf{x}_n \notin K_{n+1}$, no point of P lies in $\bigcap_{n=1}^{\infty} K_n$. Since $K_n \subset P$, this implies that $\bigcap_{n=1}^{\infty} K_n$ is empty. But each K_n is nonempty, by (iii), and $K_n \supset K_{n+1}$ by (i); this contradicts Cantor's intersection theorem (15.43). □

Corollary 15.49. *Every interval $[a,b]$ is uncountable. In particular, $\mathbb{R}$ is uncountable.*

Cantor Set

Consider the interval

$$C_0 = [0,1].$$

Remove the middle third $\left(\frac{1}{3}, \frac{2}{3}\right)$ to give

$$C_1 = \left[0, \frac{1}{3}\right] \cup \left[\frac{2}{3}, 1\right].$$

Remove the middle thirds of these intervals to give

$$C_2 = \left[0, \frac{1}{9}\right] \cup \left[\frac{2}{9}, \frac{3}{9}\right] \cup \left[\frac{6}{9}, \frac{7}{9}\right] \cup \left[\frac{8}{9}, 1\right].$$

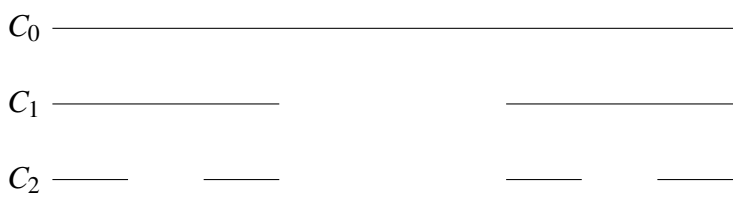


Figure 15.7: Cantor set

Repeating this process, we obtain a monotonically decreasing sequence of compact sets (C_n), where each C_n is the union of 2^n intervals, each of length 3^{-n}. Recursively,

$$C_{n+1} = \frac{1}{3}C_n \cup \left(\frac{1}{3}C_n + \frac{2}{3}\right).$$

Note that each C_n has the following properties:

(i) non-empty (since the endpoints 0 and 1 are in each C_n)

The ***Cantor set*** is defined as

$$C := \bigcap_{n=1}^{\infty} C_n.$$

The next result summarises basic properties of the Cantor set.

Lemma 15.50.

(i) C is closed.

(ii) C is compact.

(iii) C is non-empty.

(iv) C has no interior points.

Proof.

(i) C is the intersection of arbitrarily many closed sets C_n, so C is closed.

(ii) C is bounded in $[0,1]$, by definition. Since C is closed and bounded, by the Heine–Borel theorem, C is compact.

(iii) Since (C_n) is a decreasing sequence of non-empty compact sets, by Cantor's intersection theorem, $C = \bigcap_{n=1}^{\infty} C_n \neq \emptyset$.

(iv) Suppose, for a contradiction, that there exists an interior point $p \in C$. Then there exists some open interval (a,b) such that $p \in (a,b)$.

Each interval in C_n has length $\frac{1}{3^n}$. Hence for any (a,b), we can find some $n \in \mathbb{N}$ such that (a,b) is not contained in C_n, and thus not contained in C.

□

Proposition 15.51. *C is a perfect set in $\mathbb{R}$ which contains no open interval.*

Proof. We will show that (i) C contains no open interval, and (ii) C is perfect.

(i) No open interval of the form

$$\left(\frac{3k+1}{3^m}, \frac{3k+2}{3^m}\right),$$

where $k, m \in \mathbb{Z}^+$, has a point in common with C. Since every open interval (α, β) contains a open interval of the above form, if

$$\frac{1}{3^m} < \frac{\beta - \alpha}{6},$$

C contains no open interval.

(ii) Sincw we have shown that C is closed, it suffices to show that every point of C is a limit point.

Let $x \in C$, and let S be any open interval containing x. Let I_n be that interval of C_n which contains x. Choose n large enough, so that $I_n \subset S$. Let x_n be an endpoint of I_n, such that $x_n \neq x$.

It follows from the construction of C that $x_n \in C$. Hence x is a limit point of C, and C is perfect.

□

Corollary 15.52. *C is uncountable.*

Base 3 representations provide a useful way to think about the Cantor set.

Note that base 3 representations are not unique for fractions whose denominator is a power of 3; for example, $\frac{1}{3} = 0.1_3 = 0.0222\ldots_3$, where the subscript 3 denotes a base 3 representation.

- Notice that $\left(\frac{1}{3}, \frac{2}{3}\right)$ is the set of numbers in $[0,1]$ whose base 3 representations have 1 in the first digit after the decimal point. (For those numbers that have two base 3 representations, this means both such representations must have 1 in the first digit.)
- Similarly $\left(\frac{1}{9}, \frac{2}{9}\right) \cup \left(\frac{1}{3}, \frac{2}{3}\right) \cup \left(\frac{7}{9}, \frac{8}{9}\right)$ is the set of numbers in $[0,1]$ whose base 3 representations have 1 in the first digit or the second digit after the decimal point.
- And so on.
- Hence the union of all the open intervals deleted from $[0,1]$ is the set of numbers in $[0,1]$ whose base 3 representations have a 1 somewhere.

Thus we have the following description of the Cantor set.

Lemma 15.53. *C is the set of numbers in $[0,1]$ that have a base 3 representation containing only 0's and 2's.*

Remark. The phrase *a base 3 representation* indicates that if a number has two base 3 representations, then it is in the Cantor set if and only if at least one of them contains no 1s. For example, both $\frac{1}{3} = 0.0222\ldots_3 = 0.1_3$ and $\frac{2}{3} = 0.2_3 = 0.1222\ldots_3$ are in the Cantor set.

Now we can define an amazing function.

Definition 15.54 (Cantor function)**.** The Cantor function $\Lambda\colon [0,1] \to [0,1]$ is defined by converting base 3 representations into base 2 representations as follows:

- If $x \in C$, then $\Lambda(x)$ is computed from the unique base 3 representation of x containing only 0s and 2s by replacing each 2 by 1 and interpreting the resulting string as a base 2 number.
- If $x \in [0,1] \setminus C$, then $\Lambda(x)$ is computed from a base 3 representation of x by truncating after the first 1, replacing each 2 before the first 1 by 1, and interpreting the resulting string as a base 2 number.

Proposition 15.55. *The Cantor function Λ is a continuous, increasing function from $[0,1]$ onto $[0,1]$. Furthermore, $\Lambda(C) = [0,1]$.*

15.4 Connectedness

Definition 15.56 (Connectedness)**.** Let $E \subset X$. We say two disjoint non-empty sets A and B form a *separation* of E if $E = A \cup B$, and

(i) $A \cap \overline{B} = \emptyset$, and

(ii) $\overline{A} \cap B = \emptyset$;

We say E is ***connected*** if there does not exist a separation of E.

Remark. Since a connected set is defined as a negation statement, we usually prove by contradiction, or prove the contrapositive.

Remark. Separated sets are of course disjoint, but disjoint sets need not be separated. For example, $[0,1]$ and $(1,2)$ are not separated, since 1 is a limit point of $(1,2)$. However $(0,1)$ and $(1,2)$ are separated.

Example. In $\mathbb{R}^2$, the set $E = \{(x,y) \mid x,y \in \mathbb{Q}\}$ is not connected.

Proof. The sets

$$A := \{(x,y) \mid x,y \in \mathbb{Q},\, x < \sqrt{2}\}$$
$$B := \{(x,y) \mid x,y \in \mathbb{Q},\, x > \sqrt{2}\}$$

form a separation of E. □

Lemma 15.57. *Closed intervals in $\mathbb{R}$ are connected.*

Proof. Suppose, for a contradiction, that a closed interval $[a,b]$ is not connected. Then there exists non-empty disjoint sets A and B, with $A \cap \overline{B} = \emptyset$ and $\overline{A} \cap B = \emptyset$. WLOG assume $a \in A$.

Let $s = \sup A$. By 15.25, $s \in \overline{A}$. Then $\overline{A} \cap B = \emptyset$ implies $s \notin B$, so $s \in A$. Thus $A \cap \overline{B} = \emptyset$ implies $s \notin \overline{B}$. Hence there exists an open interval $(s-\varepsilon, s+\varepsilon)$ around s that is disjoint from B. But since $A \cup B = [a,b]$, we must have $(s-\varepsilon, s+\varepsilon) \subset A$. This contradicts the fact that s is the supremum of A. □

The connected subsets of $\mathbb{R}$ have a particularly simple structure:

Lemma 15.58. *$E \subset \mathbb{R}$ is connected if and only if it has the following property: if $x,y \in E$ and $x < z < y$, then $z \in E$.*

Proof.

$\boxed{\Longrightarrow}$ We prove the contrapositive.

Let $x, y \in E$ and $z \in (x, y)$ be such that $z \notin E$. Define

$$A_z := E \cap (-\infty, z), \quad B_z := E \cap (z, \infty).$$

Then $E = A_z \cup B_z$. Since $x \in A_z$ and $y \in B_z$, A_z and B_z are non-empty. Since $A_z \subset (-\infty, z)$ and $B_z \subset (z, \infty)$, they are disjoint. Hence A_z and B_z form a separation of E. Therefore E is not connected.

$\boxed{\Longleftarrow}$ We prove the contrapositive.

Suppose E is not connected. Then there are non-empty separated sets A and B such that $A \cup B = E$. Pick $x \in A$, $y \in B$, and WLOG assume $x < y$. Define

$$z := \sup(A \cap [x, y].)$$

By 15.25, $z \in \overline{A}$; hence $z \notin B$. In particular, $x \leq z < y$.

Case 1: $z \notin A$**.** It follows that $x < z < y$ and $z \notin E$.

Case 2: $z \in A$**.** Then $z \notin B$, hence there exists z_1 such that $z < z_1 < y$ and $z_1 \notin B$. Then $x < z_1 < y$ and $z_1 \notin E$.

□

Path Connectedness

15.5 Baire Category Theorem

$E \subset X$ is called ***nowhere dense*** (in X) if the interior of the closure of A is empty, i.e., $(\overline{A})^\circ = \emptyset$.

Otherwise put, E is nowhere dense iff it is contained in a closed set with empty interior. Passing to complements, we can say equivalently that E is nowhere dense iff its complement contains a dense open set (why?).

Lemma 15.59. *Let X be a metric space.*

(i) Any subset of a nowhere dense set is nowhere dense.

(ii) The union of finitely many nowhere dense sets is nowhere dense.

(iii) The closure of a nowhere dense set is nowhere dense.

(iv) If X has no isolated points, then every finite set is nowhere dense.

Proof.

(i)

(ii)

(iii)

(iv)

□

Although the union of finitely many nowhere dense sets is nowhere dense, the union of countably many nowhere dense sets need not be nowhere dense: for instance, in $X = \mathbb{R}$, the rationals $\mathbb{Q}$ are the union of countably many nowhere dense sets (why?), but the rationals are certainly not nowhere dense (indeed, they are everywhere dense, i.e. $(\overline{\mathbb{Q}})^\circ = \overline{\mathbb{Q}} = \mathbb{R}$).

This observation motivates the introduction of a larger class of sets: $A \subset X$ is called ***meager*** (or of first category) in X if it can be written as a countable union of nowhere dense sets; otherwise, it is *non-meager* (or of second category). The complement of a meager set is called ***residual***.

We then have as an immediate consequence:

Lemma 15.60. *Let X be a metric space.*

(i) Any subset of a meager set is meager.

(ii) The union of countably many meager sets is meager.

(iii) If X has no isolated points, then every countable set is meager.

We are now ready to state the Baire category theorem.

Theorem 15.61 (Baire category theorem)**.** *Let X be a complete metric space.*

(i) A meager set has empty interior.

(ii) The complement of a meager set is dense. (That is, a residual set is dense.)

(iii) A countable intersection of dense open sets is dense.

You should carefully verify that (i), (ii) and (iii) are equivalent statements, obtained by taking complements.

In applications we frequently need only the weak form of the Baire category theorem that is obtained by weakening "is dense" in (b,c) to "is non-empty" (which is valid whenever X is itself non-empty):

Corollary 15.62 (Weak form of the Baire category theorem)**.** *Let X be a non-empty complete metric space.*

(i) X cannot be written as a countable union of nowhere dense sets. (In other words, X is nonmeager in itself.)

(ii) If X is written as a countable union of closed sets, then at least one of those closed sets has nonempty interior.

(iii) A countable intersection of dense open sets is nonempty.

Exercises

Exercise 15.1. Prove that the following are metrics.

(i) On an arbitrary set X, define

$$d(x,y) = \begin{cases} 1 & (x \neq y) \\ 0 & (x = y) \end{cases}$$

(This is called the *discrete metric*.)

(ii) On $\mathbb{Z}$, define $d(x,y)$ to be 2^{-m}, where 2^m is the largest power of two dividing $x-y$. The triangle inequality holds in the following stronger form, known as the ultrametric property:

$$d(x,z) \leq \max\{d(x,y), d(y,z)\}.$$

Indeed, this is just a rephrasing of the statement that if 2^m divides both $x-y$ and $y-z$, then 2^m divides $x-z$.

(This is called the *2-adic metric*. The role of 2 can be replaced by any other prime p, and the metric may also be extended in a natural way to the rationals $\mathbb{Q}$.)

(iii) Let $\mathscr{G} = (V,E)$ be a connected graph. Define d on V as follows: $d(v,v) = 0$, and $d(v,w)$ is the length of the shortest path from v to w.

(This is known as the *path metric*.)

(iv) Let G be a group generated by elements a, b and their inverses. Define a distance on G as follows: $d(v,w)$ is the minimal k such that $v = wg_1 \cdots g_k$, where $g_i \in \{a, b, a^{-1}, b^{-1}\}$ for all i.

(This is known as the *word metric*.)

(v) Let $X = \{0,1\}^n$ (the boolean cube), the set of all strings of n zeroes and ones. Define $d(x,y)$ to be the number of coordinates in which x and y differ.

(This is known as the *Hamming distance*.)

(vi) Consider the set $P(\mathbb{R}^n)$ of one-dimensional subspaces of $\mathbb{R}^n$, that is to say lines through the origin. One way to define a distance on this set is to take, for lines L_1, L_2, the distance between L_1 and L_2 to be

$$d(L_1,L_2) = \sqrt{1 - \frac{|\langle v,w\rangle|^2}{\|v\|^2\|w\|^2}},$$

where v and w are any non-zero vectors in L_1 and L_2 respectively.

When $n = 2$, the distance between two lines is $\sin\theta$ where θ is the angle between those lines.

(This is known as the *projective space*.)

Exercise 15.2 (Product space)**.** If (X, d_X) and (Y, d_Y) are metric spaces, set

$$d_{X\times Y}\left((x_1,y_1),(x_2,y_2)\right) = \sqrt{d_X(x_1,x_2)^2 + d_Y(y_1,y_2)^2}.$$

for $x_1, x_2 \in X$, $y_1, y_2 \in Y$.
Prove that $d_{X\times Y}$ gives a metric on $X \times Y$; we call $X \times Y$ the *product space*.

Solution. Reflexivity and symmetry are obvious. Less clear is the triangle inequality. We need to prove that

$$\begin{aligned}&\sqrt{d_X(x_1,x_3)^2 + d_Y(y_1,y_3)^2} + \sqrt{d_X(x_3,x_2)^2 + d_Y(y_3,y_2)^2} \\ &\geq \sqrt{d_X(x_1,x_2)^2 + d_Y(y_1,y_2)^2}\end{aligned} \tag{1}$$

Write $a_1 = d_X(x_2,x_3)$, $a_2 = d_X(x_1,x_3)$, $a_3 = d_X(x_1,x_2)$ and similarly $b_1 = d_Y(y_2,y_3)$, $b_2 = d_Y(y_1,y_3)$ and $b_3 = d_Y(y_1,y_2)$. Thus we want to show

$$\sqrt{{a_2}^2 + {b_2}^2} + \sqrt{{a_1}^2 + {b_1}^2} \geq \sqrt{{a_3}^2 + {b_3}^2}. \tag{2}$$

To prove this, note that from the triangle inequality we have $a_1 + a_2 \geq a_3$, $b_1 + b_2 \geq b_3$. Squaring and adding gives

$${a_1}^2 + {b_1}^2 + {a_2}^2 + {b_2}^2 + 2(a_1a_2 + b_1b_2) \geq {a_3}^2 + {b_3}^2.$$

By Cauchy–Schwarz,

$$a_1a_2 + b_1b_2 \leq \sqrt{{a_1}^2 + {b_1}^2}\sqrt{{a_2}^2 + {b_2}^2}.$$

Substituting this into the previous line gives precisely the square of (2), and (1) follows. □

16 Numerical Sequences and Series

Throughout, let (X,d) be a metric space.

16.1 Convergence of Sequences

Definitions and Properties

A ***sequence*** (a_n) in X is a function $f\colon \mathbb{N} \to X$ which maps $n \mapsto a_n$.

The *range* of a sequence (a_n) is the set

$$\{x \in X \mid \exists n \in \mathbb{N}, x = a_n\}.$$

We say (a_n) is *bounded* if its range is bounded.

Definition 16.1. A sequence (a_n) ***converges*** to $a \in X$, denoted by $a_n \to a$, if

$$\forall \varepsilon > 0, \quad \exists N \in \mathbb{N}, \quad \forall n \geq N, \quad d(a_n, a) < \varepsilon. \tag{16.1}$$

We say a is a *limit* of (a_n).
If (a_n) does not converge, it is said to *diverge*.

Let us examine what (16.1) means. For any open ball centred at a, eventually all the terms of the sequence are contained in the open ball. This is illustrated in Fig. 16.1.

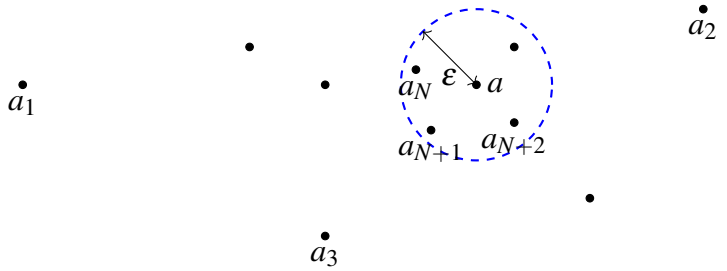


Figure 16.1: Convergence of sequence

Remark. If $a_n \not\to a$, simply negate the definition for convergence:

$$\exists \varepsilon > 0, \quad \forall N \in \mathbb{N}, \quad \exists n \geq N, \quad d(a_n, a) \geq \varepsilon.$$

Remark. From the definition, the convergence of a sequence depends not only on the sequence itself, but also on the metric space X. For instance, the sequence given by $a_n = \frac{1}{n}$ converges in $\mathbb{R}$ (to 0), but fails to converge in $\mathbb{R}^+$. In cases of possible ambiguity, we shall specify "convergent in X" rather than "convergent".

Lemma. *If a sequence converges, its limit is unique.*

Proof. Suppose $a_n \to a$ and $a_n \to a'$ for $a, a' \in X$. We will show that $a = a'$.

Let $\varepsilon > 0$ be given. There exist $N, N' \in \mathbb{N}$ such that

$$n \geq N \implies d(a_n, a) < \frac{\varepsilon}{2}$$
$$n \geq N' \implies d(a_n, a') < \frac{\varepsilon}{2}$$

Take $N_1 := \max\{N, N'\}$. If $n \geq N_1$, then both hold. By the triangle inequality,

$$d(a, a') \leq d(a, a_n) + d(a_n, a') < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Since this holds for all $\varepsilon > 0$, we must have $d(a, a') = 0$. Hence $a = a'$. □

Since the limit is unique, we can give it a notation.

Notation. If (a_n) converges to a, we denote $\lim_{n\to\infty} a_n = a$.

Example. $\lim_{n\to\infty} \frac{1}{n} = 0$.

Proof. Let $\varepsilon > 0$ be given. By the Archimedean property, there exists $N \in \mathbb{N}$ such that $\frac{1}{N} < \varepsilon$. Take $N = \left\lfloor \frac{1}{\varepsilon} \right\rfloor + 1$. Then for all $n \geq N$,

$$\left| \frac{1}{n} - 0 \right| = \frac{1}{n} \leq \frac{1}{N} = \frac{1}{\left\lfloor \frac{1}{\varepsilon} \right\rfloor + 1} < \frac{1}{\frac{1}{\varepsilon}} = \varepsilon$$

as desired. Therefore $\frac{1}{n} \to 0$. □

A useful tip for finding the required N (in terms of ε) is to *work backwards* from the result we wish to show, as illustrated in the following example.

Example. $\lim_{n\to\infty} \left(1 + (-1)^n \frac{1}{\sqrt{n}}\right) = 1$.

Let $a_n = 1 + (-1)^n \frac{1}{\sqrt{n}}$. Before our proof, we need to find $N \in \mathbb{N}$ such that if $n \geq N$,

$$\begin{aligned} & |a_n - 1| < \varepsilon \\ \iff & \left|(-1)^n \frac{1}{\sqrt{n}}\right| = \frac{1}{\sqrt{n}} < \varepsilon \\ \iff & \frac{1}{n} < \varepsilon^2 \\ \iff & n > \frac{1}{\varepsilon^2} \end{aligned}$$

Hence take $N = \left\lfloor \frac{1}{\varepsilon^2} \right\rfloor + 1$.

Proof. Let $\varepsilon > 0$ be given. Take $N = \left\lfloor \frac{1}{\varepsilon^2} \right\rfloor + 1$. If $n \geq N$, then

$$\begin{aligned} |a_n - 1| &= \left|(-1)^n \frac{1}{\sqrt{n}}\right| = \frac{1}{\sqrt{n}} \\ &\leq \frac{1}{\sqrt{N}} = \frac{1}{\sqrt{\left\lfloor \frac{1}{\varepsilon^2} \right\rfloor + 1}} \\ &< \frac{1}{\sqrt{\frac{1}{\varepsilon^2}}} = \varepsilon \end{aligned}$$

as desired. Therefore $a_n \to 1$. □

We outline some important properties of convergent sequences.

Lemma 16.2. *Let (a_n) be a sequence in X.*

(i) $a_n \to a$ if and only if every open ball of a contains a_n for all but finitely many n.

(ii) Every convergent sequence is bounded.

(iii) Suppose $E \subset X$. Then a is a limit point of E if and only if there exists a sequence (a_n) in $E \setminus \{a\}$ such that $a_n \to a$.

Proof.

(i) $\boxed{\Longrightarrow}$ Suppose $a_n \to a$. Let $\varepsilon > 0$ be given, there exists $N \in \mathbb{N}$ such that

$$n \geq N \implies d(a_n, a) < \varepsilon \implies a_n \in B_\varepsilon(a).$$

$\boxed{\Longleftarrow}$ Suppose every open ball of a contains all but finitely many of the a_n.

Let $\varepsilon > 0$ be given. Consider the open ball $B_\varepsilon(a)$. Since $B_\varepsilon(a)$ is a open ball of a, it

will also eventually contain all a_n; that is, there exists $N \in \mathbb{N}$ such that if $n \geq N$, then $a_n \in B_\varepsilon(a)$, i.e. $d(a_n, a) < \varepsilon$. Hence $a_n \to a$.

(ii) Suppose $a_n \to a$. Then (16.1) holds. In particular, if we take $\varepsilon = 1$, there exists $N \in \mathbb{N}$ such that

$$n \geq N \implies d(a_n, a) < 1.$$

Let

$$r = \max\{1, d(a_1, a), \ldots, d(a_N, a)\}.$$

Then $d(a_n, a) \leq r$ for all $n \in \mathbb{N}$, so the range of a_n is bounded by $B_r(a)$. Hence (a_n) is bounded.

(iii) $\boxed{\implies}$ Consider a sequence of open balls $\left(B_{\frac{1}{n}}(a)\right)$. Since a is a limit point of E, each open ball intersects with E at some point which is not a. Pick one such point a_n from each $B_{\frac{1}{n}}(a) \cap E$. Then

$$d(a_n, a) < \frac{1}{n}.$$

Let $\varepsilon > 0$ be given. By the Archimedean property, there exists $N \in \mathbb{N}$ such that $\frac{1}{N} < \varepsilon$. If $n \geq N$,

$$d(a_n, a) < \frac{1}{n} \leq \frac{1}{N} < \varepsilon,$$

which shows that $a_n \to a$.

$\boxed{\impliedby}$ Suppose there exists a sequence (a_n) in $E \setminus \{a\}$ such that $a_n \to a$. By (i), for each open ball $B_\varepsilon(a)$, there exists $N \in \mathbb{N}$ such that if $n \geq N$,

$$a_n \in B_\varepsilon(a).$$

Since $a_n \in E \setminus \{a\}$, this shows that a is a limit point of E.

□

A useful consequence of (ii) is its contrapositive: any unbounded sequence is divergent.

The converse of (ii) is not true. **Counterexample:** $(-1)^n$.

Remark. The limit of a sequence need not be a limit point. For instance, consider a constant sequence whose terms are all equal to an isolated point.

Lemma 16.3 (Ordering). *Suppose (a_n) and (b_n) are convergent sequences, and $a_n \leq b_n$. Then*

$$\lim_{n\to\infty} a_n \leq \lim_{n\to\infty} b_n.$$

Proof. Let $a = \lim_{n\to\infty} a_n$, $b = \lim_{n\to\infty} b_n$. Suppose, for a contradiction, that $a > b$.

Let $\varepsilon = a - b > 0$ be given. There exists $N_1, N_2 \in \mathbb{N}$ such that

$$n \geq N_1 \implies |a_n - a| < \frac{\varepsilon}{2},$$
$$n \geq N_2 \implies |b_n - b| < \frac{\varepsilon}{2}.$$

Let $N = \max\{N_1, N_2\}$. Then $n \geq N$ implies

$$a_n > a - \frac{\varepsilon}{2}, \quad b_n < b + \frac{\varepsilon}{2}$$

and thus

$$a_n - b_n > a - b - \varepsilon = 0$$

so $a_n > b_n$, which is a contradiction. □

Remark. If $a_n < b_n$, we may not necessarily have $\lim_{n\to\infty} a_n < \lim_{n\to\infty} b_n$. For instance, $-\frac{1}{n} < \frac{1}{n}$ but their limits are both 0.

The next result shows that limits preserve the familiar arithmetic properties.

Lemma 16.4 (Limit laws). *Suppose (a_n) and (b_n) are convergent seqeunces in $\mathbb{C}$; let $a = \lim_{n\to\infty} a_n$, $b = \lim_{n\to\infty} b_n$. Then*

(i) $\lim_{n\to\infty} ca_n = ca$, *where c is a constant* *(scalar multiplication)*

(ii) $\lim_{n\to\infty} (a_n + b_n) = a + b$ *(addition)*

(iii) $\lim_{n\to\infty} (a_n b_n) = ab$ *(multiplication)*

(iv) $\lim_{n\to\infty} \frac{a_n}{b_n} = \frac{a}{b}$ $(b_n \neq 0,\ b \neq 0)$ *(division)*

Proof.

(i) The case where $c = 0$ is trivial. Now suppose $c \neq 0$. Let $\varepsilon > 0$ be given. Then there exists $N \in \mathbb{N}$ such that

$$n \geq N \implies |a_n - a| < \frac{\varepsilon}{|c|}.$$

Then if $n \geq N$,

$$|ca_n - ca| = |c|\,|a_n - a| < \varepsilon.$$

(ii) Let $\varepsilon > 0$ be given. There exist $N_1, N_2 \in \mathbb{N}$ such that

$$n \geq N_1 \implies |a_n - a| < \frac{\varepsilon}{2},$$
$$n \geq N_2 \implies |b_n - b| < \frac{\varepsilon}{2}.$$

Let $N = \max\{N_1, N_2\}$. If $n \geq N$,

$$\begin{aligned} |(a_n + b_n) - (a + b)| &\leq |a_n - a| + |b_n - b| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

(iii) Write

$$a_n b_n - ab = (a_n - a)(b_n - b) + a(b_n - b) + b(a_n - a).$$

Let $\varepsilon > 0$ be given. There exist $N_1, N_2 \in \mathbb{N}$ such that

$$\begin{aligned} n \geq N_1 &\implies |a_n - a| < \sqrt{\varepsilon}, \\ n \geq N_2 &\implies |b_n - b| < \sqrt{\varepsilon}. \end{aligned}$$

Let $N = \max\{N_1, N_2\}$. If $n \geq N$,

$$|(a_n - a)(b_n - b)| < \varepsilon,$$

and thus $\lim_{n\to\infty} (a_n - a)(b_n - b) = 0$.

Note that $\lim_{n\to\infty} a(b_n - b) = \lim_{n\to\infty} b(a_n - a) = 0$. Hence

$$\lim_{n\to\infty} (a_n b_n - ab) = 0.$$

(iv) Since we have proven multiplication in (iii), it suffices to show that $\lim_{n\to\infty} \frac{1}{b_n} = \frac{1}{b}$.

Let $\varepsilon > 0$. There exists $m \in \mathbb{N}$ such that

$$n \geq m \implies |b_n - b| < \frac{1}{2}|b|.$$

There exists $N \in \mathbb{N}$, $N > m$ such that

$$n \geq N \implies |b_n - b| < \frac{1}{2}|b|^2 \varepsilon.$$

Hence if $n \geq N$,

$$\left| \frac{1}{b_n} - \frac{1}{b} \right| = \left| \frac{b - b_n}{b_n b} \right| < \frac{2}{|b|^2} |b_n - b| < \varepsilon.$$

□

We now prove the analogue for Euclidean spaces.

Lemma 16.5.

(i) Suppose $(\mathbf{x}_n)$ is a sequence in $\mathbb{R}^k$, where

$$\mathbf{x}_n = (x_{1,n}, \dots, x_{k,n}) \quad (n = 1, 2, \dots).$$

Let $\mathbf{x} = (x_1, \dots, x_k)$. Then $\mathbf{x}_n \to \mathbf{x}$ if and only if

$$\lim_{n\to\infty} x_{i,n} = x_i \quad (1 \le i \le k).$$

(ii) Suppose $(\mathbf{x}_n)$ and $(\mathbf{y}_n)$ are sequences in $\mathbb{R}^k$, (β_n) is a sequence of real numbers, and $\mathbf{x}_n \to \mathbf{x}$, $\mathbf{y}_n \to \mathbf{y}$, $\beta_n \to \beta$. Then

$$\lim_{n\to\infty} (\mathbf{x}_n + \mathbf{y}_n) = \mathbf{x} + \mathbf{y}, \quad \lim_{n\to\infty} \mathbf{x}_n \cdot \mathbf{y}_n = \mathbf{x} \cdot \mathbf{y}, \quad \lim_{n\to\infty} \beta_n \mathbf{x}_n = \beta \mathbf{x}.$$

Proof.

(i) $\boxed{\Longrightarrow}$ Let $i \in \{1, \dots, k\}$. The definition of the norm in $\mathbb{R}^k$ implies

$$\begin{aligned} \|\mathbf{x}_n - \mathbf{x}\| &= \left(|x_{1,n} - x_i|^2 + \cdots + |x_{i,n} - x_i|^2 + \cdots + |x_{k,n} - x_k|^2\right)^{1/2} \\ &\ge \left(|x_{i,n} - x_i|^2\right)^{1/2} \\ &= |x_{i,n} - x_i|. \end{aligned}$$

Let $\varepsilon > 0$ be given. Since $\mathbf{x}_n \to \mathbf{x}$, there exists $N \in \mathbb{N}$ such that for all $n \ge N$,

$$\|\mathbf{x}_n - \mathbf{x}\| < \varepsilon.$$

This implies $|x_{i,n} - x_i| < \varepsilon$. Hence $\lim_{n\to\infty} x_{i,n} = x_i$.

$\boxed{\Longleftarrow}$ Suppose $\lim_{n\to\infty} x_{i,n} = x_i$ for $i = 1, \dots, k$. Then for each $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that $n \ge N$ implies

$$|x_{i,n} - x_i| < \frac{\varepsilon}{\sqrt{k}} \quad (i = 1, \dots, k).$$

Hence $n \ge N$ implies

$$\|\mathbf{x}_n - \mathbf{x}\| = \left(\sum_{i=1}^{k} |x_{i,n} - x_i|^2\right)^{1/2} < \varepsilon,$$

so that $\mathbf{x}_n \to \mathbf{x}$.

(ii) This follows from (i) and 16.4.

□

The next result provides a useful method to evaluate limits of sequences.

Lemma 16.6 (Squeeze theorem). *Let $a_n \le c_n \le b_n$ where (a_n) and (b_n) converge and $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = L$. Then (c_n) converges, and*

$$\lim_{n\to\infty} c_n = L.$$

Proof. Let $\varepsilon > 0$ be given. There exist $N_1, N_2 \in \mathbb{N}$ such that

$$n \ge N_1 \implies |a_n - L| < \varepsilon,$$
$$n \ge N_2 \implies |b_n - L| < \varepsilon.$$

In particular, we have

$$a_n > L - \varepsilon, \quad b_n < L + \varepsilon.$$

Let $N = \max\{N_1, N_2\}$. If $n \ge N$,

$$L - \varepsilon < a_n \le c_n \le b_n < L + \varepsilon$$

or

$$|c_n - L| < \varepsilon.$$

Hence (c_n) converges, and $c_n \to L$. □

Example. $\lim_{n\to\infty} \dfrac{\sin n}{n} = 0.$

Proof. We have $-1 \le \sin n \le 1$, so

$$-\frac{1}{n} \le \frac{\sin n}{n} \le \frac{1}{n}.$$

Now

$$\lim_{n\to\infty} \frac{1}{n} = \lim_{n\to\infty} \left(-\frac{1}{n}\right) = 0,$$

so the squeeze theorem yields the desired result. □

Subsequences

Definition 16.7 (Subsequence). Given a sequence (a_n), consider a sequence (n_k) of positive integers such that $n_1 < n_2 < \cdots$. Then (a_{n_k}) is called a ***subsequence*** of (a_n). If (a_{n_k}) converges, its limit is called a *subsequential limit* of (a_n).

Example. The (divergent) subsequence $1, \frac{1}{2}, 1, \frac{1}{3}, 1, \frac{1}{4}, \ldots$ has, among its subsequences, the sequence $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$ as well as the sequence $1, \frac{1}{4}, 1, \frac{1}{9}, 1, \frac{1}{16}, \ldots$.

The next result states that subsequences of a convergent sequence must converge to the same limit as the sequence.

Lemma 16.8. *(a_n) converges to a if and only if every subsequence of (a_n) converges to a.*

Proof.

$\boxed{\Longrightarrow}$ Suppose $a_n \to a$. Let $\varepsilon > 0$ be given. Then there exists $N \in \mathbb{N}$ such that

$$n \geq N \implies d(a_n, a) < \varepsilon.$$

Every subsequence of (a_n) can be written in the form (a_{n_k}) where $n_1 < n_2 < \cdots$ is a strictly increasing sequence of positive integers. Pick M such that $n_M \geq N$. Then

$$k > M \implies n_k > n_M \geq N \implies d(a_{n_k}, a) < \varepsilon.$$

Hence every subsequence of (a_n) converges to a.

$\boxed{\Longleftarrow}$ Suppose every subsequence of (a_n) converges to a. Since (a_n) is a subsequence of itself, $a_n \to a$. □

Corollary 16.9. *If (a_n) has a divergent subsequence, then (a_n) diverges.*

Example. Define the sequence

$$a_n = \begin{cases} \frac{1}{n} & (n \text{ is odd}) \\ n & (n \text{ is even}) \end{cases}$$

The even terms form the subsequence $(a_{2n}) = 2, 4, 6, 8, \ldots$ which diverges. Hence (a_n) diverges.

Corollary 16.10. *If (a_n) has convergent subsequences whose limits are not equal, then (a_n) diverges.*

Compactness of a metric space guarantees the existence of a convergent subsequence.

Proposition 16.11. *In a compact metric space, any sequence has a convergent subsequence.*

Proof. Suppose (a_n) is a sequence in a compact metric space X.

Let E be the range of (a_n). We consider two cases:

Case 1: E is finite. There are infinitely many terms in the sequence (a_n), but only finitely many distinct terms in E. By the pigeonhole principle, at least one term of E appears infinitely many times in the sequence.

That is, there exists $a \in E$ and a sequence (n_k) with $n_1 < n_2 < \cdots$ such that

$$a_{n_1} = a_{n_2} = \cdots = a.$$

This subsequence (a_{n_k}) evidently converges to a.

Case 2: E is infinite. If E is infinite, then E is an infinite subset of a compact set. By 15.34, E has a limit point $a \in X$.

We now construct a subsequence (a_{n_k}) of (a_n) such that $a_{n_k} \to a$.

- Choose n_1 so that $d(a, a_{n_1}) < 1$.
- Having chosen $n_1, \ldots, n_{k-1}$, choose n_k where $n_k > n_{k-1}$ such that $d(a, a_{n_k}) < \frac{1}{k}$ (such n_k exists due to 15.20).

Then $a_{n_k} \to a$.

□

Corollary 16.12 (Bolzano–Weierstrass)**.** *Every bounded sequence in $\mathbb{R}^k$ has a convergent subsequence.*

Proof. By 15.39, every bounded sequence in $\mathbb{R}^k$ lives in a compact subset of $\mathbb{R}^k$, and therefore it lives in a compact metric space.

Hence by 16.11, it contains a convergent subsequence converging to a point in $\mathbb{R}^k$. □

Proposition 16.13. *Suppose (a_n) is a sequence in X. Then the subsequential limits of (a_n) form a closed subset of X.*

Proof. Let E^* be the set of all subsequential limits of (a_n). Let $q \in X$ be a limit point of E^*. We will show $q \in E^*$.

To show q is a subsequential limit, we will construct a subsequence (a_{n_k}) which converges to q.

Step 0: If $E^* = \emptyset$, then E^* is a closed subset of X.

Step 1: Choose n_1 so that $a_{n_1} \neq q$. (If no such n_1 exists, then $E^* = \{q\}$ has only one point, so E^* is a closed subset of X.) Let $\delta = d(q, a_{n_1})$. (Then $\delta > 0$.)

Step k: Suppose $n_1, \ldots, n_{k-1}$ are chosen. Since q is a limit point of E^*, there exists $a \in E^*$ such that $d(a,q) < \frac{1}{2^k}\delta$.

Since $a \in E^*$, a is a subsequential limit of (a_n), so there exists $n_k > n_{k-1}$ such that $d(a, a_{n_k}) < \frac{1}{2^k}\delta$.

Thus for $k = 1, 2, \ldots,$

$$\begin{aligned} d(q, a_{n_k}) &\leq d(q,a) + d(a, a_{n_k}) \\ &< \frac{1}{2^k}\delta + \frac{1}{2^k}\delta = \frac{1}{2^{k-1}}\delta. \end{aligned}$$

This implies (a_{n_k}) converges to q. Hence $q \in E^*$. □

Cauchy Sequences

This is a very helpful way to determine whether a sequence is convergent or divergent, as it does not require the limit to be known. Subsequently we will see many instances where the convergence of all sorts of limits are compared with similar counterparts; generally we describe such properties as *Cauchy criteria*.

Definition 16.14 (Cauchy sequence)**.** A sequence (a_n) in X is a ***Cauchy sequence*** if

$$\forall \varepsilon > 0, \quad \exists N \in \mathbb{N}, \quad \forall n, m \geq N, \quad d(a_n, a_m) < \varepsilon.$$

Remark. Intuitively, the distances between any two terms becomes sufficiently small after a certain point.

Lemma 16.15. *In any metric space, every convergent sequence is a Cauchy sequence.*

Proof. Suppose $a_n \to a$. Let $\varepsilon > 0$. There exists $N \in \mathbb{N}$ such that for all $n \geq N$,

$$d(a_n, a) < \frac{\varepsilon}{2}.$$

Then for all $n, m \geq N$,

$$d(a_n, a_m) \leq d(a_n, a) + d(a_m, a) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Hence (a_n) is a Cauchy sequence. □

The converse is not true. **Counterexample:** the sequence $3, 3.1, 3.14, 3.141, 3.1415, \ldots$ is a Cauchy sequence but does not converge in $\mathbb{Q}$.

Example. The sequence (a_n) defined by

$$a_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}$$

does not converge in $\mathbb{R}$.

Proof. We claim that (a_n) is not a Cauchy sequence. WLOG assume $n > m$. Consider

$$|a_n - a_m| = \frac{1}{m+1} + \frac{1}{m+2} + \cdots + \frac{1}{n} \geq \frac{n-m}{n} = 1 - \frac{m}{n}.$$

Let $n = 2m$, then

$$|a_n - a_m| = |a_{2m} - a_m| > \frac{1}{2}.$$

Hence (a_n) is not a Cauchy sequence, so it does not converge. □

We have a special name for metric spaces where the converse is true.

Definition 16.16. A metric space is ***complete*** if every Cauchy sequence converges.

We now show several examples of metric spaces that are complete.

Proposition 16.17.

(i) Compact metric spaces are complete.

(ii) Euclidean spaces $\mathbb{R}^k$ are complete.

Proof.

(i) Suppose X is a metric space. Let (a_n) be a Cauchy sequence in X. Since X is compact, it is sequentially compact. Then there exists a subsequence (a_{n_k}) such that $a_{n_k} \to a$.

Claim. $a_n \to a$.

Let $\varepsilon > 0$. Since (a_n) is a Cauchy sequence, there exists $N_1 \in \mathbb{N}$ such that

$$n, m \geq N_1 \implies d(a_n - a_m) < \frac{\varepsilon}{2}.$$

$a_{n_k} \to a$ implies there exists $N_2 \in \mathbb{N}$ such that

$$n_k \geq N_2 \implies d(a_{n_k}, a) < \frac{\varepsilon}{2}.$$

Let $N = \max\{N_1, N_2\}$, fix some $n_k \geq N$. Then $n \geq N$ implies

$$d(a_n, a) \leq d(a_n, a_{n_k}) + d(a_{n_k}, a) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

(ii) Suppose (a_n) is a Cauchy sequence.

We perform three steps:

(a) We first show that (a_n) is bounded:
Pick $N \in \mathbb{N}$ such that $|a_n - a_N| \leq 1$ for all $n \geq N$. Then

$$|a_n| \leq \max\{1 + |a_N|, |a_1|, \ldots, |a_{N-1}|\}.$$

(b) Since (a_n) is bounded, by Bolzano–Weierstrass, (a_n) contains a subsequence (a_{n_k}) which converges to a.

(c) We now show that $a_n \to a$.
Let $\varepsilon > 0$ be given. Since (a_n) is a Cauchy sequence, there exists $N_1 \in \mathbb{N}$ such that

$$n, m \geq N_1 \implies |a_n - a_m| < \frac{\varepsilon}{2}.$$

Since $a_{n_k} \to a$, there exists $M \in \mathbb{N}$ such that for all $k > M$,

$$n_k > n_M \implies |a_{n_k} - a| < \frac{\varepsilon}{2}.$$

Now since $n_1 < n_2 < \cdots$ is a sequence of strictly increasing positive integers, we can pick $i > M$ such that $n_k > N_1$. Then for all $n \geq N_1$, by setting $m = n_k$ we obtain

$$|a_n - a_{n_k}| < \frac{\varepsilon}{2}, \quad |a_{n_k} - a| < \frac{\varepsilon}{2}.$$

Hence

$$|a_n - a| \leq |a_n - a_{n_k}| + |a_{n_k} - a| < \varepsilon.$$

Therefore (a_n) is convergent, and $a_n \to a$.

□

This implies that every closed subset E of a complete metric space X is complete. (Every Cauchy sequence in E is a Cauchy sequence in X, hence it converges to some $a \in X$, and actually $a \in E$ since E is closed.)

Monotonic Sequences

Definition 16.18 (Monotonic sequence)**.** A sequence (a_n) in $\mathbb{R}$ is

(i) *monotonically increasing* if $a_n \leq a_{n+1}$ for $n \in \mathbb{N}$;

(ii) *monotonically decreasing* if $a_n \geq a_{n+1}$ for $n \in \mathbb{N}$;

(iii) ***monotonic*** if it is either monotonically increasing or monotonically decreasing.

For monotonic sequences in $\mathbb{R}$, convergence is equivalent to boundedness.

Lemma 16.19 (Monotone convergence theorem)**.** *A monotonic real sequence converges if and only if it is bounded.*

Proof. We show the case for monotically increasing sequences; the case for monotonically decreasing sequences is similar.

$\boxed{\Longrightarrow}$ We already proved that a convergent sequence is bounded.

$\boxed{\Longleftarrow}$ Suppose (a_n) is a monotonically increasing sequence that is bounded above.

Let E be the range of (a_n). Then E is bounded above; let $a = \sup E$.

Claim. $a_n \to a$.

By definition of supremum, $a_n \leq a$ for all $n \in \mathbb{N}$. For every $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that

$$a - \varepsilon < a_N \leq a,$$

otherwise $a - \varepsilon$ would be an upper bound of E. Since (a_n) is monotically increasing, $n \geq N$ implies $a_N \leq a_n \leq a$, so

$$a - \varepsilon < a_n \leq a,$$

which implies $|a_n - a| < \varepsilon$. Hence $a_n \to a$. □

Lemma 16.20 (Monotone subsequence theorem)**.** *Every real sequence has a monotone subsequence.*

Proof. Suppose (a_n) is a real sequence. We say a_m is a *peak* if $a_n \leq a_m$ for all $n > m$.

Case 1: (a_n) **has infinitely many peaks.** List them: $a_{m_1}, a_{m_2}, \ldots$. This is a monotically decreasing subsequence of (a_n).

Case 2: (a_n) **has finitely many peaks.** Let a_{n_1} be past the last peak. This point is not a peak, so there exists $n_2 > n_1$ such that $a_{n_2} > a_{n_1}$.

But a_{n_2} is not a peak either, so there exists $n_3 > n_2$ such that $a_{n_3} > a_{n_2}$.

Continuing in this inductively gives a monotonically increasing subsequence (a_{n_k}) of (a_n).

□

We present an alternative proof for Bolzano–Weierstrass theorem, in the case of $\mathbb{R}$:

Every bounded real sequence has a convergent subsequence.

Proof. Let (a_n) be a bounded real sequence.

By the monotone subsequence theorem, (a_n) has a monotone subsequence (a_{n_k}).

Since (a_n) is bounded, so is (a_{n_k}).

By the monotone convergence theorem, (a_{n_k}) converges. □

Properly Divergent Sequences

We define *properly divergent sequences* as the following:

Definition 16.21. Suppose (a_n) is a sequence in $\mathbb{R}$. We write $a_n \to \infty$ if

$$\forall M \in \mathbb{R}, \quad \exists N \in \mathbb{N}, \quad \forall n \geq N, \quad a_n \geq M.$$

Similarly, we write $a_n \to -\infty$ if

$$\forall M \in \mathbb{R}, \quad \exists N \in \mathbb{N}, \quad \forall n \geq N, \quad a_n \leq M.$$

We also write

$$\lim_{n\to\infty} a_n = +\infty, \quad \lim_{n\to\infty} a_n = -\infty.$$

Remark. $+\infty, -\infty$ are symbols, not real numbers.

Example. If (a_n) is increasing and unbounded, then $a_n \to +\infty$.

Proof. Let $M \in \mathbb{R}$. Since (a_n) is unbounded, there exists $N \in \mathbb{N}$ such that

$$x_N > M.$$

Since (a_n) is increasing, $a_n \geq a_N$ for all $n \geq N$. Thus

$$\forall n \geq N, \quad a_n > M.$$

□

The following are special cases of the above:

- $\lim_{n\to\infty} n = +\infty$.
- $\lim_{n\to\infty} n^k = +\infty$ where $k \in \mathbb{N}$.
- $\lim_{n\to\infty} b^n = +\infty$ where $b > 1$.
- $\lim_{n\to\infty} \left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right) = +\infty$.

Limit Superior and Inferior

Definition 16.22. Let (a_n) be a real sequence. Let $E \subset \overline{\mathbb{R}}$ be the set of subsequential limits (possibly including $+\infty, -\infty$). Define respectively the ***limit superior*** and ***limit infimum*** of (a_n) as

$$\limsup_{n\to\infty} a_n := \sup E,$$
$$\liminf_{n\to\infty} a_n := \inf E.$$

Intuitively, limsup is the largest value that subsequences can approach; liminf is the smallest value that subsequences can approach.

Remark. The limit superior and limit infimum exist due to the existence of supremum and infimum in $\overline{\mathbb{R}}$. If (a_n) is not bounded above, then some subsequence tends to $+\infty$, so

$$\limsup_{n\to\infty} a_n = \infty.$$

Similarly, if (a_n) is not bounded below, then some subsequence tends to $-\infty$, so

$$\liminf_{n\to\infty} a_n = -\infty.$$

Example.

- Let (a_n) be a sequence containing all rationals. Then every real number is a subsequantial limit, so
$$\limsup_{n\to\infty} a_n = +\infty, \quad \liminf_{n\to\infty} a_n = -\infty.$$
- Let $a_n = \dfrac{(-1)^n}{1+\frac{1}{n}}$. When n is even, $a_n = \frac{1}{1+\frac{1}{n}} \to 1$; when n is odd, $a_n = \frac{-1}{1+\frac{1}{n}} \to -1$. Thus
$$\limsup_{n\to\infty} a_n = 1, \quad \liminf_{n\to\infty} a_n = -1.$$
- For a real-valued sequence (a_n), $\lim\limits_{n\to\infty} a_n = a$ if and only if
$$\limsup_{n\to\infty} a_n = \liminf_{n\to\infty} a_n = a.$$
This follows since all subsequences must converge to the same limit as the original sequence.

The next result is a useful characterisation of limsup.

Lemma 16.23. *Suppose (a_n) is a real sequence. Then*

(i) $\limsup_{n\to\infty} a_n \in E$;

(ii) if $a > \limsup_{n\to\infty} a_n$, *there exists* $N \in \mathbb{N}$ *such that* $a_n < a$ *for all* $n \geq N$.

Moreover, $\limsup_{n\to\infty} a_n$ *is the only number that satisfies (i) and (ii).*

Proof.

(i) We consider three cases for the value of $\limsup_{n\to\infty} a_n$:

Case 1: $\limsup_{n\to\infty} a_n = +\infty$**.** Then $\sup E = +\infty$, so E is not bounded above. Hence (a_n) is not bounded above, so (a_n) has a subsequence (a_{n_k}) such that $a_{n_k} \to \infty$. Thus $\limsup_{n\to\infty} a_n = +\infty \in E$.

Case 2: $\limsup_{n\to\infty} a_n \in \mathbb{R}$**.** Then $\sup E \in \mathbb{R}$, so E is bounded above. Hence at least one subsequential limit exists, so that (i) follows from 16.13 and Theorem 2.28.

Case 3: $\limsup_{n\to\infty} a_n = -\infty$**.** Then $\sup E = -\infty$, so E contains only one element, namely $-\infty$. Hence (a_n) has no subsequential limit. Thus for any $M \in \mathbb{R}$, $a_n > M$ for at most a finite number of values of n, so that $a_n \to -\infty$.

(ii) We prove by contradiction.

Suppose there is a number $a > \limsup_{n\to\infty} a_n$ such that $a_n \geq a$ for infinitely many values of n. In that case, there is a number $y \in E$ such that $y \geq a > \limsup_{n\to\infty} a_n$, contradicting the definition of $\limsup_{n\to\infty} a_n$.

We now show uniqueness. Suppose, for a contradiction, that two numbers p and q satisfy (i) and (ii). WLOG assume $p < q$. Then choose a such that $p < a < q$. Since p satisfies (i), we have $a_n < a$ for all $n \geq N$. But then q cannot satisfy (i). □

Of course, an analogous result is true for $\liminf_{n\to\infty} a_n$.

Lemma 16.24 (Comparison)**.** *If $a_n \leq b_n$ for $n \geq N$ (where N is fixed), then*

$$\liminf_{n\to\infty} a_n \leq \liminf_{n\to\infty} b_n,$$
$$\limsup_{n\to\infty} a_n \leq \limsup_{n\to\infty} b_n.$$

Lemma 16.25.

$$\liminf_{n\to\infty} a_n = -\limsup_{n\to\infty}(-a_n).$$

Proof. Exercise; use the definitions and 14.11. □

Lemma 16.26 (Arithmetic properties)**.**

(i) *If* $k > 0$, $\limsup_{n\to\infty} ka_n = k\limsup_{n\to\infty} a_n$.
If $k < 0$, $\limsup_{n\to\infty} ka_n = k\liminf_{n\to\infty} a_n$.

(ii)

$$\limsup_{n\to\infty}(a_n + b_n) \le \limsup_{n\to\infty} a_n + \limsup_{n\to\infty} b_n$$

(iii)

$$\limsup_{n\to\infty}(a_n + b_n) \ge \limsup_{n\to\infty} a_n + \liminf_{n\to\infty} b_n$$

(iv) *For any constant* C,

$$\limsup_{n\to\infty}(C + a_n) = C + \limsup_{n\to\infty} a_n.$$

Proof.

$$\begin{aligned}\limsup_{n\to\infty} a_n &= \limsup_{n\to\infty}(a_n + b_n - b_n) \le \limsup_{n\to\infty}(a_n + b_n) + \limsup_{n\to\infty}(-b_n) \\ &= \limsup_{n\to\infty}(a_n + b_n) - \liminf_{n\to\infty} b_n.\end{aligned}$$

□

16.2 More on Sequences

o-notation and Big o-notation

The o-notation and O-notation are used to compare the size of some given sequence relative to some well known sequence.

The o-notation is, roughly speaking, used to compare two sequences when one is much smaller than the other. x_n is said to be much smaller than y_n, denoted $x_n \ll y_n$, if $\lim_{n\to\infty} \frac{x_n}{y_n} = 0$, denoted by $x_n = o(y_n)$. We usually use this notation when both sequences approach 0, for example $\frac{1}{n^2} = o\left(\frac{1}{n}\right)$.

The O-notation tries to measure how fast a sequence grows or shrinks We denote $x_n = O(y_n)$ if there exists a constant $M > 0$ and a natural number $N \in \mathbb{N}$ such that $|x_n| \leq My_n$ for all $n > N$. For example $2n^2 \sin n = O(n^2)$.

Example. Let (r_k) be a sequence of all rational numbers strictly between 0 and 1 where each rational number appears exactly once. Show that $\limsup r_k = 1$ and $\liminf r_k = 0$.

16.3 Convergence of Series

Given a sequence (a_n), we associate a sequence (s_n), where

$$s_n = \sum_{k=1}^{n} a_k = a_1 + a_2 + \cdots + a_n,$$

where the term s_n is called the *n-th partial sum*. The sequence (s_n) is often written as

$$\sum_{n=1}^{\infty} a_n,$$

which we call a ***series***.

Definition 16.27 (Convergence of series)**.** We say the series *converges* if $s_n \to s$ (the sequence of partial sums converges), and write $\sum_{n=1}^{\infty} a_n = s$; that is,

$$\forall \varepsilon > 0, \quad \exists N \in \mathbb{N}, \quad \forall n \geq N, \quad \left| \sum_{k=1}^{n} a_k - s \right| < \varepsilon.$$

The number s is called the *sum* of the series. If (s_n) diverges, the series is said to *diverge*.

Notation. When there is no possible ambiguity, we write $\sum_{n=1}^{\infty} a_n$ simply as $\sum a_n$.

The Cauchy criterion can be restated in the following form:

Lemma 16.28 (Cauchy criterion)**.** $\sum a_n$ *converges if and only if*

$$\forall \varepsilon > 0, \quad \exists N \in \mathbb{N}, \quad \forall n \geq m \geq N, \quad \left| \sum_{k=m}^{n} a_k \right| \leq \varepsilon.$$

Convergence Tests

To determine the convergence of a series, apart from using the definition and the Cauchy criterion, we also have the following methods:

- Divergence test (16.29)
- Boundedness of partial sums (16.30, for series of non-negative terms)
- Comparison test (16.31)

- Root test (16.32)
- Ratio test (16.33)
- Absolute convergence (16.35)

Lemma 16.29 (Divergence test). *If $a_n \not\to 0$, then $\sum a_n$ diverges.*

Proof. We prove the contrapositive: if $\sum a_n$ converges, then $a_n \to 0$.

In the Cauchy criterion, take $m = n$, then $|a_n| \leq \varepsilon$ for all $n \geq N$. □

The converse is not true. **Counterexample:** harmonic series.

Lemma 16.30. *A series of non-negative terms converges if and only if its partial sums form a bounded sequence.*

Proof. Partial sums are monotonically increasing. But bounded monotonic sequences converge. □

Lemma 16.31 (Comparison test). *Consider two sequences (a_n) and (b_n).*

(i) Suppose $|a_n| \leq b_n$ for all $n \geq N_0$ (where N_0 is some fixed integer). If $\sum b_n$ converges, then $\sum a_n$ converges.

(ii) Suppose $a_n \geq b_n \geq 0$ for all $n \geq N_0$. If $\sum b_n$ diverges, then $\sum a_n$ diverges.

Proof.

(i) Let $\varepsilon > 0$ be given. Since $\sum b_n$ converges, by the Cauchy criterion, there exists $N \in \mathbb{N}$, $N \geq N_0$ such that for $n \geq m \geq N$,

$$\sum_{k=m}^{n} b_k \leq \varepsilon.$$

By the triangle inequality,

$$\left| \sum_{k=m}^{n} a_k \right| \leq \sum_{k=m}^{n} |a_k| \leq \sum_{k=m}^{n} b_k \leq \varepsilon.$$

By the Cauchy criterion, $\sum a_n$ converges.

(ii) We prove the contrapositive. Suppose $\sum a_n$ converges. Since $|b_n| \leq a_n$ for all $n \geq N_0$, by (i), $\sum b_n$ converges.

□

To employ the comparison test, we need to be familiar with several series whose convergence or divergence is known.

Example (Geometric series)**.** A geometric series takes the form

$$\sum_{n=0}^{\infty} x^n.$$

Lemma.

(i) If $|x| < 1$*, then* $\sum x^n$ *converges, and*

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}.$$

(ii) If $|x| \geq 1$*, then* $\sum x^n$ *diverges.*

Proof.

(i) For $|x| < 1$, the n-th partial sum is

$$\sum_{k=0}^{n} x^k = 1 + x + x^2 + \cdots + x^n. \tag{1}$$

Multiplying both sides of (1) by x gives

$$x \sum_{k=0}^{n} x^k = x + x^2 + x^3 \cdots + x^{n+1}. \tag{2}$$

Taking the difference of (1) and (2),

$$(1-x) \sum_{k=0}^{n} x^k = 1 - x^{n+1}$$

so

$$\sum_{k=0}^{n} x^k = \frac{1-x^{n+1}}{1-x}.$$

Taking limits $n \to \infty$ yields the desired result.

(ii) For $|x| \geq 1$, $x^n \not\to 0$. By the divergence test, $\sum x^n$ diverges.

□

Example (p-series)**.** A p-series takes the form

$$\sum_{n=1}^{\infty} \frac{1}{n^p}.$$

To determine the convergence of p-series, we first prove the following lemma, which states that a rather "thin" subsequence of (a_n) determines the convergence of $\sum a_n$.

Lemma (Cauchy condensation test)**.** *Suppose $a_1 \geq a_2 \geq \cdots \geq 0$. Then $\sum a_n$ converges if and only if the series*

$$\sum_{k=0}^{\infty} 2^k a_{2^k} = a_1 + 2a_2 + 4a_4 + \cdots$$

converges.

Proof. Let s_n and t_k denote the n-th partial sum of (a_n) and the k-th partial sum of $(2^k a_{2^k})$ respectively; that is,

$$\begin{aligned} s_n &= a_1 + a_2 + \cdots + a_n, \\ t_k &= a_1 + 2a_2 + \cdots + 2^k a_{2^k}. \end{aligned}$$

We consider two cases:

- For $n < 2^k$, group terms to give

$$\begin{aligned} s_n &= a_1 + a_2 + \cdots + a_n \\ &\leq a_1 + (a_2 + a_3) + \cdots + (a_{2^k} + \cdots + a_{2^{k+1}-1}) \\ &\leq a_1 + 2a_2 + \cdots + 2^k a_{2^k} \\ &= t_k. \end{aligned}$$

  By comparison test, if (t_k) converges, then (s_n) converges.

- For $n > 2^k$,

$$\begin{aligned} s_n &\geq a_1 + a_2 + (a_3 + a_4) + \cdots + (a_{2^{k-1}+1} + \cdots + a_{2^k}) \\ &\geq \frac{1}{2} a_1 + a_2 + 2a_4 + \cdots + 2^{k-1} a_{2^k} \\ &= \frac{1}{2} t_k. \end{aligned}$$

  By comparison test, if (s_n) converges, then (t_k) converges.

□

Proposition (p-test)**.**

(i) If $p > 1$, $\sum \frac{1}{n^p}$ converges.

(ii) If $p \leq 1$, $\sum \frac{1}{n^p}$ diverges.

Proof. Note that if $p \leq 0$, then $\frac{1}{n^p} \not\to 0$. By the divergence test, $\sum \frac{1}{n^p}$ diverges.
If $p > 0$, we want to apply the above lemma. Consider the series

$$\sum_{k=0}^{\infty} 2^k \cdot \frac{1}{(2^k)^p} = \sum_{k=0}^{\infty} 2^{(1-p)k} = \sum_{k=0}^{\infty} \left(2^{1-p}\right)^k,$$

which is a geometric series. Hence the above series converges if and only if $|2^{1-p}| < 1$, which holds if and only if $1 - p < 0$. Then apply the above lemma to conclude the convergence of $\frac{1}{n^p}$. □

Remark. If $p = 1$, the resulting series is known as the *harmonic series* (which diverges). If $p = 2$, the resulting series converges, and the sum of this series is $\frac{\pi^2}{6}$ (Basel problem).

Example (The number e)**.** Consider the series

$$\sum_{n=0}^{\infty} \frac{1}{n!}.$$

We first show that the above series converges. Consider the n-th partial sum:

$$\begin{aligned}\sum_{k=0}^{n} \frac{1}{k!} &= \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{n!} \\ &\leq 1 + 1 + \frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^{n-1}} \\ &< 1 + 1 + \frac{1}{2} + \frac{1}{2^2} + \cdots = 3.\end{aligned}$$

Since the partial sums are bounded (by 3), and the terms are non-negative, the series converges. Then we can make the following definition for the sum of the series:

$$e := \sum_{n=0}^{\infty} \frac{1}{n!}$$

Proposition. *e is irrational.*

Proof. Suppose, for a contradiction, that e is rational. Then $e = \frac{p}{q}$, where p and q are positive integers. Let s_n denote the n-th partial sum:

$$s_n = \sum_{k=0}^{n} \frac{1}{k!}.$$

Then

$$\begin{aligned} e - s_n &= \frac{1}{(n+1)!} + \frac{1}{(n+2)!} + \frac{1}{(n+3)!} + \cdots \\ &< \frac{1}{(n+1)!}\left(1 + \frac{1}{n+1} + \frac{1}{(n+1)^2} + \cdots\right) \\ &= \frac{1}{(n+1)!} \cdot \frac{n+1}{n} = \frac{1}{n!n} \end{aligned}$$

and thus

$$0 < e - s_n < \frac{1}{n!n}.$$

Taking $n = q$ and multiplying both sides by $q!$ gives

$$0 < q!(e - s_q) < \frac{1}{q}.$$

Note that $q!e$ is an integer (by assumption), and

$$q!s_q = q!\left(1 + 1 + \frac{1}{2!} + \cdots + \frac{1}{q!}\right)$$

is an integer, so $q!(e - s_q)$ is an integer. Since $q \geq 1$, this implies the existence of an integer between 0 and 1, which is absurd. Hence we have reached a contradiction. □

Lemma. $\lim_{n\to\infty}\left(1 + \frac{1}{n}\right)^n = e.$

Proof. Let

$$s_n = \sum_{k=0}^{n} \frac{1}{k!}, \quad t_n = \left(1 + \frac{1}{n}\right)^n.$$

By the binomial theorem,

$$t_n = 1 + 1 + \frac{1}{2!}\left(1 - \frac{1}{n}\right) + \frac{1}{3!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) + \cdots + \frac{1}{n!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{n-1}{n}\right).$$

Comparing term by term, $t_n \leq s_n$. Thus by 16.24,

$$\limsup_{n\to\infty} t_n \leq \limsup_{n\to\infty} s_n = e.$$

Next, if $n \geq m$,

$$t_n \geq 1 + 1 + \frac{1}{2!}\left(1 - \frac{1}{n}\right) + \cdots + \frac{1}{m!}\left(1 - \frac{1}{n}\right)\cdots\left(1 - \frac{m-1}{n}\right).$$

Let $n \to \infty$, keeping m fixed. We get

$$\liminf_{n\to\infty} t_n \geq 1 + 1 + \frac{1}{2!} + \cdots + \frac{1}{m!},$$

so that

$$s_m \leq \liminf_{n\to\infty} t_n.$$

Letting $m \to \infty$, we get

$$e \leq \liminf_{n\to\infty} t_n.$$

Thus it follows that

$$\limsup_{n\to\infty} t_n = \liminf_{n\to\infty} t_n = e,$$

so the desired result follows. □

Lemma 16.32 (Root test). *Given $\sum a_n$, let $\alpha = \limsup_{n\to\infty} \sqrt[n]{|a_n|}$.*

(i) If $\alpha < 1$, $\sum a_n$ converges.

(ii) If $\alpha > 1$, $\sum a_n$ diverges.

(iii) If $\alpha = 1$, the test gives no information.

Remark. We use limsup since the limsup of a sequence always exists (in $\overline{\mathbb{R}}$), while the limit may not necessarily exist.

Proof.

(i) Suppose $\alpha < 1$. Choose β such that $\alpha < \beta < 1$. Since $\beta > \limsup_{n\to\infty} \sqrt[n]{|a_n|}$, by 16.23, there exists $N \in \mathbb{N}$ such that for all $n \geq N$,

$$\sqrt[n]{|a_n|} < \beta.$$

Thus

$$|a_n| < \beta^n.$$

Since $0 < \beta < 1$, the geometric series $\sum \beta^n$ converges. By the comparison test, $\sum a_n$ converges.

(ii) Suppose $\alpha > 1$. By 16.23, α is a subsequential limit of the sequence $(\sqrt[n]{|a_n|})$. Thus there exists a subsequence $(\sqrt[n_k]{|a_{n_k}|})$ such that

$$\sqrt[n_k]{|a_{n_k}|} \to \alpha.$$

Thus $|a_n| > 1$ for infinitely many values of n. Hence $a_n \not\to 0$. By the divergence test, $\sum a_n$ diverges.

(iii) Consider the series

$$\sum \frac{1}{n} \quad \text{and} \quad \sum \frac{1}{n^2}.$$

For each of these series $\alpha = 1$, but the first diverges, the second converges. Hence the condition that $\alpha = 1$ does not give us information on the convergence of a series.

□

Lemma 16.33 (Ratio test). *The series* $\sum a_n$

(i) converges if $\limsup_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| < 1$;

(ii) diverges if $\left|\frac{a_{n+1}}{a_n}\right| \geq 1$ *for all* $n \geq N_0$ *(where* N_0 *is some fixed integer).*

Proof.

(i) If $\limsup_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| < 1$, there exists $\beta < 1$ and $N \in \mathbb{N}$ such tht for all $n \geq N$,

$$\left|\frac{a_{n+1}}{a_n}\right| < \beta.$$

In particular, from $n = N$ to $n = N + k$,

$$\begin{aligned} |a_{N+1}| &< \beta |a_N| \\ |a_{N+2}| &< \beta |a_{N+1}| < \beta^2 |a_N| \\ &\vdots \\ |a_{N+k}| &< \beta^k |a_N| \end{aligned}$$

Hence for all $n \geq N$,

$$\begin{aligned} |a_n| &< |a_N| \beta^{n-N} \\ &= \left(|a_N| \beta^{-N}\right) \beta^n \end{aligned}$$

Taking the sum on both sides gives

$$\sum |a_n| < |a_N| \beta^{-N} \sum \beta^n.$$

Since $\beta < 1$, the geometric series $\sum \beta^n$ converges. By the comparison test, $\sum a_n$ converges.

(ii) Suppose $\left|\frac{a_{n+1}}{a_n}\right| \geq 1$ for all $n \geq N_0$. Then $|a_{n+1}| \geq |a_n|$ for $n \geq N_0$, so $a_n \not\to 0$. By the divergence test, $\sum a_n$ diverges.

□

The ratio test is easier to apply than the root test (since it is usually easier to compute ratios than n-th roots), but the root test is more powerful:

Proposition 16.34. *For any sequence (a_n) of positive numbers,*

$$\liminf_{n\to\infty} \frac{a_{n+1}}{a_n} \le \liminf_{n\to\infty} \sqrt[n]{a_n},$$
$$\limsup_{n\to\infty} \sqrt[n]{a_n} \le \limsup_{n\to\infty} \frac{a_{n+1}}{a_n}.$$

Proof. We shall prove the second inequality; the proof of the first is similar. Let

$$\alpha = \limsup_{n\to\infty} \frac{a_{n+1}}{a_n}.$$

If $\alpha = +\infty$, there is nothing to prove. If α is finite, choose $\beta > \alpha$. Then there exists $N \in \mathbb{N}$ such that for all $n \ge N$,

$$\frac{a_{n+1}}{a_n} \le \beta.$$

In particular, for any $p > 0$,

$$a_{N+k+1} \le \beta a_{N+k} \quad (k = 0, 1, \ldots, p-1).$$

Multiplying these inequalities, we obtain

$$a_{N+p} \le \beta^p a_N,$$

or

$$a_n \le a_N \beta^{-N} \cdot \beta^n \quad (n \ge N).$$

Hence

$$\sqrt[n]{a_n} \le \sqrt[n]{a_N \beta^{-N}} \cdot \beta,$$

so that

$$\limsup_{n\to\infty} \sqrt[n]{a_n} \le \beta$$

by 16.23. Since this is true for every $\beta > \alpha$, we have

$$\limsup_{n\to\infty} \sqrt[n]{a_n} \le \alpha.$$

□

The series $\sum a_n$ is said to *converge absolutely* if the series $\sum |a_n|$ converges.

The next result shows that absolute convergence implies convergence.

Lemma 16.35. *If $\sum a_n$ converges absolutely, then $\sum a_n$ converges.*

Proof. Suppose $\sum a_n$ converges absolutely. Then $\sum |a_n|$ converges. Let $\varepsilon > 0$ be given. By the Cauchy criterion, there exists $N \in \mathbb{N}$ such that for all $n \geq m \geq N$,

$$\left| \sum_{k=m}^{n} |a_k| \right| < \varepsilon.$$

Since all the terms are non-negative,

$$\sum_{k=m}^{n} |a_k| < \varepsilon.$$

By the triangle inequality,

$$\left| \sum_{k=m}^{n} a_k \right| \leq \sum_{k=m}^{n} |a_k| < \varepsilon.$$

Hence by the Cauchy criterion, $\sum a_n$ converges. □

Note that the converse may not necessarily be true. We say that $\sum a_n$ is *conditionally convergent* if it converges, but does not converge absolutely.

Example. The alternating harmonic series defined by

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}$$

converges to $\ln 2$, but it is not absolutely convergent (since the harmonic series diverges).

Summation by Parts

Proposition 16.36 (Partial summation formula). *Given two sequences (a_n) and (b_n), let the n-partial sum of (a_n) be denoted by*

$$A_n = \sum_{k=0}^{n} a_k$$

for $n \geq 0$; let $A_{-1} = 0$. Then, if $0 \leq p \leq q$,

$$\sum_{n=p}^{q} a_n b_n = \sum_{n=p}^{q-1} A_n (b_n - b_{n+1}) + A_q b_q - A_{p-1} b_p.$$

Proof. The RHS can be written as

$$\begin{aligned}
&\sum_{n=p}^{q-1} A_n b_n + A_q b_q - \sum_{n=p}^{q-1} A_n b_{n+1} - A_{p-1} b_p \\
&= \sum_{n=p}^{q} A_n b_n - \sum_{n=p-1}^{q-1} A_n b_{n+1} \\
&= \sum_{n=p}^{q} A_n b_n - \sum_{n=p}^{q} A_{n-1} b_n \\
&= \sum_{n=p}^{q} (A_n - A_{n-1})\, b_n \\
&= \sum_{n=p}^{q} a_n b_n
\end{aligned}$$

which is equal to the LHS. □

Suppose that we have a series $\sum a_n b_n$ and we wish to show that it converges, then there are these two strategies:

Lemma 16.37 (Dirichlet's test). *Suppose (a_n) and (b_n) are sequences such that*

- *the partial sums A_n of $\sum a_n$ form a bounded sequence,*
- *$b_0 \geq b_1 \geq b_2 \geq \cdots$,*
- *$b_n \to 0$.*

Then $\sum a_n b_n = 0$.

Proof. Since the partial sums A_n form a bounded sequence, there exists M such that

$$|A_n| \leq M \quad (\forall n \in \mathbb{N})$$

Since $b_n \to 0$, fix $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that

$$b_N \leq \frac{\varepsilon}{2M}.$$

For $q \geq p \geq N$, by the partial summation formula, we have

$$\begin{aligned}
\left| \sum_{n=p}^{q} a_n b_n \right| &= \left| \sum_{n=p}^{q-1} A_n (b_n - b_{n+1}) + A_q b_q - A_{p-1} b_p \right| \\
&\leq M \left| \sum_{n=p}^{q-1} (b_n - b_{n+1}) + b_q + b_p \right| \quad [\because |A_n| \leq M] \\
&= M \left| (b_p - b_q) + b_q + b_p \right| = 2M b_p \leq 2M b_n \leq \varepsilon.
\end{aligned}$$

By the Cauchy criterion, $\sum a_n b_n$ converges to 0. □

The following is a convenient application of Dirichlet's test.

Corollary 16.38 (Alternating series test). *Suppose (c_n) is a sequence such that*

- $|c_1| \geq |c_2| \geq |c_3| \geq \cdots$,
- $c_{2m-1} \geq 0, c_{2m} \leq 0$ *for* $m = 1, 2, 3, \ldots$,
- $c_n \to 0$.

Then $\sum c_n = 0$.

Proof. Let

$$a_n = (-1)^{n+1}, \quad b_n = |c_n|.$$

Note that

- the partial sums of (a_n) are 0s and 1s, so they are bounded;
- $b_0 \geq b_1 \geq b_2 \geq \cdots$ holds by assumption;
- $c_n \to 0$ implies $|c_n| \to 0$, so $b_n \to 0$.

Then by 16.37, we have that $\sum a_n b_n = 0$, so $\sum c_n = 0$. □

Lemma 16.39 (Abel's test). *If $\sum a_n$ converges, and (b_n) is monotonic and bounded, then $\sum a_n b_n$ converges.*

Proof. Suppose that $\sum a_n$ is convergent, and (b_n) is monotonic and bounded. Now what we do is that we try to transform b_n into the one that we see in Dirichlet's test.

Since (b_n) is monotonic, it's either monotonically increasing or decreasing. If (b_n) were to be monotonically increasing, we multiply the sequences (a_n) and (b_n) by -1; $\sum a_n$ is still convergent here. Thus assume (b_n) is monotonically decreasing.

Since (b_n) is monotonically decreasing, the sequence (b_n) has an infimum which is also its limit, so we may let $b = \lim b_n = \inf b_n$. Then $(b_n - b)$ is a monotonically decreasing sequence which converges to 0

Since $\sum a_n$ is convergent, its partial sums A_n is bounded

By Dirichlet's test, this shows that $\sum a_n(b_n - b)$ is convergent

However, we have the additional requirement that $\sum a_n$ is convergent, so we can easily deal with the extra '$-b$' term just by noting that $\sum a_n b$ is convergent and hence

$$\sum a_n b_n = \sum a_n(b_n - b) + \sum a_n b$$

is convergent. □

16.4 More on Series

Addition and Multiplication of Series

Lemma 16.40. *If $\sum a_n = A$ and $\sum b_n = B$, then*

(i) $\sum(a_n + b_n) = A + B$, *(addition)*

(ii) $\sum ca_n = cA$ *for some constant* c. *(scalar multiplication)*

Proof.

(i) Let the n-th partial sums be denoted by

$$A_n = \sum_{k=0}^{n} a_k, \quad B_n = \sum_{k=0}^{n} b_k.$$

Then

$$\lim_{n\to\infty} \sum_{k=0}^{n} (a_k + b_k) = \lim_{n\to\infty} (A_n + B_n) = \lim_{n\to\infty} A_n + \lim_{n\to\infty} B_n = A + B.$$

(ii) Simply factor out the constant c:

$$\lim_{n\to\infty} \sum_{k=0}^{n} ca_k = c \lim_{n\to\infty} \sum_{k=0}^{n} a_k = cA.$$

□

The situation becomes more complicated when we consider multiplication of two series. To begin with, we have to define the product. This can be done in several ways; we shall consider the so-called "Cauchy product".

Definition 16.41 (Cauchy product). Given $\sum a_n$ and $\sum b_n$, let

$$c_n = \sum_{k=0}^{n} a_k b_{n-k} \quad (n = 0, 1, 2, \dots)$$

We call $\sum c_n$ the *product* of the two given series.

This definition may be motivated as follows. If we take two power series $\sum a_n z^n$ and $\sum b_n z^n$,

multiply them term by term, and collect terms containing the same power of z, we get

$$\begin{aligned}\left(\sum_{n=0}^{\infty} a_n z^n\right)\left(\sum_{n=0}^{\infty} b_n z^n\right) &= \left(a_0 + a_1 z + a_2 z^2 + \cdots\right)\left(b_0 + b_1 z + b_2 z^2 + \cdots\right) \\ &= a_0 b_0 + (a_0 b_1 + a_1 b_0) z + (a_0 b_2 + a_1 b_1 + a_2 b_0) z^2 + \cdots \\ &= c_0 + c_1 z + c_2 z^2 + \cdots\end{aligned}$$

Setting $z = 1$, we arrive at the above definition.

Note that $\sum c_n$ may not converge, even if $\sum a_n$ and $\sum b_n$ do. However $\sum c_n$ converges if an additional condition is imposed: at least one of the two series converges absolutely.

Proposition 16.42 (Mertens' theorem). *Suppose $\sum a_n = A$, $\sum b_n = B$, and $\sum a_n$ converges absolutely. Then their Cauchy product converges to AB.*

Proof. Let $\sum c_n$ be the Cauchy product of $\sum a_n$ and $\sum b_n$. Let the n-th partial sums be denoted by

$$A_n = \sum_{k=0}^{n} a_k, \quad B_n = \sum_{k=0}^{n} b_k, \quad C_n = \sum_{k=0}^{n} c_k.$$

Also let $\beta_n = B_n - B$. Then

$$\begin{aligned}C_n &= a_0 b_0 + (a_0 b_1 + a_1 b_0) + \cdots + (a_0 b_n + a_1 b_{n-1} + \cdots + a_n b_0) \\ &= a_0 B_n + a_1 B_{n-1} + \cdots + a_n B_0 \\ &= a_0 (B + \beta_n) + a_1 (B + \beta_{n-1}) + \cdots + a_n (B + \beta_0) \\ &= A_n B + (a_0 \beta_n + a_1 \beta_{n-1} + \cdots + a_n \beta_0)\end{aligned}$$

Our goal is to show that $C_n \to AB$. Since $A_n B \to AB$, it suffices to show that

$$\gamma_n = a_0 \beta_n + a_1 \beta_{n-1} + \cdots + a_n \beta_0 \to 0.$$

We now use the absolute convergence of (a_n); let $\alpha = \sum |a_n|$. Fix $\varepsilon > 0$, there exists $N_1 \in \mathbb{N}$ such that

$$n \geq N_1 \implies \sum_{k=0}^{n} |a_k| - \alpha < \varepsilon$$

since the terms are non-negative. Since $B_n \to B$, $\beta_n \to 0$. Then there exists $N_2 \in \mathbb{N}$ such that

$$n \geq N_2 \implies |\beta_n| \leq \varepsilon.$$

Let $N = \max\{N_1, N_2\}$. Then for $n \geq N$, by triangle inequality,

$$\begin{aligned}|\gamma_n| &= |\beta_0 a_n + \cdots + \beta_n a_0| \\ &\leq |\beta_0 a_n + \cdots + \beta_N a_{n-N}| + |\beta_{N+1} a_{n-N-1} + \cdots + \beta_n a_0| \\ &\leq |\beta_0 a_n + \cdots + \beta_N a_{n-N}| + \varepsilon(|a_{n-N-1}| + \cdots + |a_0|) \\ &\leq |\beta_0 a_n + \cdots + \beta_N a_{n-N}| + \varepsilon\alpha.\end{aligned}$$

Keeping N fixed, and letting $n \to \infty$, we get

$$\limsup_{n\to\infty} |\gamma_n| \leq \varepsilon\alpha,$$

sine $a_n \to 0$. Since ε is arbitrary, we have $\gamma_n \to 0$, as desired. □

Proposition 16.43 (Abel's theorem). *Let $\sum a_n = A$, $\sum b_n = B$, $\sum c_n = C$, where $\sum c_n$ is the Cauchy product of $\sum a_n$ and $\sum b_n$. Then $C = AB$.*

Rearrangements

Definition 16.44 (Rearrangement). Let (k_n) be a sequence in which every positive integer appears once and only once. Let

$$a'_n = a_{k_n} \quad (\forall n \in \mathbb{N})$$

We say that $\sum a'_n$ is a *rearrangement* of $\sum a_n$.

If (s_n) and (s'_n) are the sequences of partial sums of (a_n) and (a'_n) respectively, it is easily seen that, in general, these two sequences consist of entirely different numbers. We are thus led to the problem of determining under what conditions all rearrangements of a convergent series will converge and whether the sums are necessarily the same.

Theorem 16.45 (Riemann series theorem). *Let $\sum a_n$ be a series of real numbers which converges, but not absolutely. Suppose $-\infty \leq \alpha \leq \beta \leq \infty$. Then there exists a rearrangement $\sum a'_n$ with partial sums s'_n such that*

$$\liminf_{n\to\infty} s'_n = \alpha, \quad \limsup_{n\to\infty} s'_n = \beta.$$

Proof. Let

$$p_n = \frac{|a_n| + a_n}{2}, \quad q_n = \frac{|a_n| - a_n}{2} \quad (n = 1, 2, \ldots).$$

Then $p_n - q_n = a_n$, $p_n + q_n = |a_n|$, $p_n \geq 0$, $q_n \geq 0$.

Claim. The series $\sum p_n$ and $\sum q_n$ must both diverge.

If both were convergent, then

$$\sum (p_n + q_n) = \sum |a_n|$$

would converge, contrary to hypothesis. Since

$$\sum_{n=1}^{N} a_n = \sum_{n=1}^{N} (p_n - q_n) = \sum_{n=1}^{N} p_n - \sum_{n=1}^{N} q_n,$$

divergence of $\sum p_n$ and convergence of $\sum q_n$ (or vice versa) implies divergence of $\sum a_n$, again contrary to hypothesis.

Now let $P_1, P_2, \ldots$ denote the non-negative terms of $\sum a_n$, in the order which they occur, and let $Q_1, Q_2, \ldots$ be the absolute values of the negative terms of $\sum a_n$, also in their original order.

The series $\sum P_n$ and $\sum Q_n$ differ from $\sum p_n$ and $\sum q_n$ only by zero terms, and are therefore divergent.

We shall construct sequences (m_n) and (k_n), such that the series

$$\begin{aligned} &(P_1 + \cdots + P_{m_1}) - (Q_1 + \cdots + Q_{k_1}) + \\ &(P_{m_1+1} + \cdots + P_{m_2}) - (Q_{k_1+1} + \cdots + Q_{k_2}) + \cdots \end{aligned} \tag{1}$$

which clearly is a rearrangement of $\sum a_n$, satisfies $\liminf_{n\to\infty} s'_n = \alpha$, $\limsup_{n\to\infty} s'_n = \beta$.

Choose real-valued sequences (α_n) and (β_n) such that $\alpha_n \to \alpha$, $\beta_n \to \beta$, $\alpha_n < \beta_n$, $\beta_1 > 0$.

Let m_1, k_1 be the smallest integers such that

$$\begin{aligned} P_1 + \cdots + P_{m_1} &> \beta_1, \\ P_1 + \cdots + P_{m_1} - (Q_1 + \cdots + Q_{k_1}) &< \alpha_1; \end{aligned}$$

let m_2, k_2 be the smallest integers such that

$$\begin{aligned} (P_1 + \cdots + P_{m_1}) - (Q_1 + \cdots + Q_{k_1}) + (P_{m_1+1} + \cdots + P_{m_2}) &> \beta_2 \\ (P_1 + \cdots + P_{m_1}) - (Q_1 + \cdots + Q_{k_1}) + (P_{m_1+1} + \cdots + P_{m_2}) - (Q_{k_1+1} + \cdots + Q_{k_2}) &< \alpha_2; \end{aligned}$$

and continue in this way. This is possible since $\sum P_n$ and $\sum Q_n$ diverge.

If x_n, y_n denote the partial sums of (1) whose last terms are P_{m_n}, $-Q_{k_n}$, then

$$|x_n - \beta_n| \le P_{m_n}, \quad |y_n - \alpha_n| \le Q_{k_n}.$$

Since $P_n \to 0$ and $Q_n \to 0$ as $n \to \infty$, wee see that $x_n \to \beta$, $y_n \to \alpha$.

Finally, it is clear that no number less than α or greater than β can be a subsequential limit of the partial sums of (1).

to review

□

Theorem 16.46. *If $\sum a_n$ is a series of complex numbers which converges absolutely, then every rearrangement of $\sum a_n$ converges, and they all converge to the same sum.*

Proof. Let $\sum a'_n$ be a rearrangement, with partial sums s'_n. Since $\sum a_n$ converges absolutely, given $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that $m \geq n \geq N$ implies

$$\sum_{i=n}^{m} |a_i| < \varepsilon. \tag{1}$$

Now choose p such that the integers $1, 2, \ldots, N$ are all contained in the set $k_1, \ldots, k_p$ (we use the notation of Definition 16.44). Then if $n > p$, the numbers $a_1, \ldots, a_N$ will cancel in the difference $s_n - s'_n$, so that $|s_n - s'_n| < \varepsilon$, by (1) Hence (s'_n) converges to the same sum as (s_n). □ to review

Exercises

Exercise 16.1. Show the following:

(i) $\lim_{n\to\infty} \frac{1}{n^p} = 0$ $(p > 0)$

(ii) $\lim_{n\to\infty} \sqrt[n]{p} = 1$ $(p > 0)$

(iii) $\lim_{n\to\infty} \sqrt[n]{n} = 1$

(iv) $\lim_{n\to\infty} \frac{n^\alpha}{(1+p)^n} = 0$ $(p > 0,\ \alpha \in \mathbb{R})$

(v) $\lim_{n\to\infty} x^n = 0$ $(|x| < 1)$

Solution.

(i) Let $\varepsilon > 0$ be given. By the Archimedean property, there exists N^p such that $\frac{1}{N^p} < \varepsilon$. Then $N > \left(\frac{1}{\varepsilon}\right)^{\frac{1}{p}}$. Thus $n \geq N$ implies

$$\left|\frac{1}{n^p} - 0\right| = \frac{1}{n^p} \leq \frac{1}{N^p} < \varepsilon.$$

(ii) Consider different values of p.

Case 1: $p > 1$**.** Let $a_n = \sqrt[n]{p} - 1$. We will show $a_n \to 0$.

Note that $a_n > 0$. By the binomial theorem,

$$1 + na_n \leq (1 + a_n)^n = p,$$

so that

$$0 < a_n \leq \frac{p-1}{n}.$$

By the squeeze theorem, $a_n \to 0$.

Case 2: $p = 1$**.** Trivial.

Case 3: $0 < p < 1$**.** Let $p = \frac{1}{q}$ for some $q > 1$. Using Case 1, taking reciprocals yields

$$\lim_{n\to\infty} \sqrt[n]{p} = \lim_{n\to\infty} \sqrt[n]{\frac{1}{q}} = \frac{1}{\lim_{n\to\infty} \sqrt[n]{q}} = 1.$$

(iii) Let $a_n = \sqrt[n]{n} - 1$. Then $a_n \geq 0$, and, by the binomial theorem,

$$n = (1 + a_n)^n \geq \frac{n(n-1)}{2} {a_n}^2.$$

Hence

$$0 \le a_n \le \sqrt{\frac{2}{n-1}} \quad (n \ge 2.)$$

By the squeeze theorem, $a_n \to 0$.

(iv) Let k be an integer such that $k > \alpha$, $k > 0$. For $n > 2k$,

$$(1+p)^n > \binom{n}{k} p^k = \frac{n(n-1)\cdots(n-k+1)}{k!} p^k > \frac{n^k p^k}{2^k k!}.$$

Hence

$$0 < \frac{n^\alpha}{(1+p)^n} < \frac{2^k k!}{p^k} n^{\alpha-k} \quad (n > 2k).$$

Since $\alpha - k < 0$, by (i), $n^{\alpha-k} \to 0$.

(v) Take $\alpha = 0$ in (iv).

□

The first few exercises involve evaluating limits. The general approach involves algebraic manipulation, inequalities, the limit laws, and other results to convert the original limit problem into another one that has been already understood and solved.

Exercise 16.2. Evaluate $\lim_{n\to\infty} \dfrac{5n^2-3n+7}{6n^2+n+2}$.

Solution. Since $\frac{1}{n} \to 0$,

$$\frac{5n^2-3n+7}{6n^2+n+2} = \frac{5-\frac{3}{n}+\frac{7}{n^2}}{6+\frac{1}{n}+\frac{2}{n^2}} \to \frac{5-0+0}{6+0+0} = \frac{5}{6}.$$

□

Exercise 16.3. Evaluate $\lim_{n\to\infty} \sqrt{n+1} - \sqrt{n}$.

Solution. Write

$$\begin{aligned}\sqrt{n+1} - \sqrt{n} &= \frac{\left(\sqrt{n+1}-\sqrt{n}\right)\left(\sqrt{n+1}-\sqrt{n}\right)}{\sqrt{n+1}+\sqrt{n}} \\ &= \frac{1}{\sqrt{n+1}+\sqrt{n}} < \frac{1}{\sqrt{n}}.\end{aligned}$$

Since $\lim_{n\to\infty} \dfrac{1}{\sqrt{n}} = 0$, by the squeeze theorem, we conclude that $\lim_{n\to\infty} \sqrt{n+1} - \sqrt{n} = 0$. □

Exercise 16.4. Show that $\displaystyle\int_1^\infty \frac{\cos t}{t^2}\,\mathrm{d}t$ converges.

Solution. Consider the sequence $a_n = \displaystyle\int_1^n \frac{\cos t}{t^2}\,\mathrm{d}t$. We will show that (a_n) is a Cauchy sequence and thus converges.

It is easy to see that (a_n) is bounded, for

$$|a_n| = \left|\int_1^n \frac{\cos t}{t^2}\,\mathrm{d}t\right| \le \int_1^n \frac{|\cos t|}{t^2}\,\mathrm{d}t \le \int_1^n \frac{1}{t^2}\,\mathrm{d}t = \left[-\frac{1}{t}\right]_1^n = 1 - \frac{1}{n} \le 1$$

for all $n \in \mathbb{N}$.

Fix $\varepsilon > 0$. Pick $N > \frac{1}{\varepsilon}$. For $n, m \ge N$ with $n > m$,

$$a_n - a_m = \int_1^n \frac{\cos t}{t^2}\,\mathrm{d}t - \int_1^m \frac{\cos t}{t^2}\,\mathrm{d}t = \int_m^n \frac{\cos t}{t^2}\,\mathrm{d}t\,.$$

Hence

$$|a_n - a_m| = \left|\int_m^n \frac{\cos t}{t^2}\,\mathrm{d}t\right| \le \int_m^n \frac{|\cos t|}{t^2}\,\mathrm{d}t \le \int_m^n \frac{1}{t^2}\,\mathrm{d}t = \left[-\frac{1}{t}\right]_m^n = \frac{1}{m} - \frac{1}{n} < \frac{1}{m} \le \frac{1}{N} < \varepsilon.$$

Therefore the sequence (a_n) converges. □

Exercise 16.5. Let (x_n) be a real sequence, let $\alpha \ge 2$ be a constant. Define the sequence (y_n) as follows:

$$y_n = x_n + \alpha x_{n+1} \quad (n = 1, 2, \dots)$$

Show that if (y_n) is convergent, then (x_n) is also convergent.

Exercise 16.6 ([Rud76] 3.1)**.** Prove that the convergence of (a_n) implies the convergence of $(|a_n|)$. Is the converse true?

Solution. Let $\varepsilon > 0$ be given. Since (a_n) is a Cauchy sequence, there exists $N \in \mathbb{N}$ such that for all $n, m \ge N$,

$$|a_n - a_m| < \varepsilon.$$

See that

$$||a_n| - |a_m|| \le |a_n - a_m| < \varepsilon,$$

so $(|a_n|)$ is a Cauchy sequence, and therefore must converge.

The converse is not true, as shown by the sequence (a_n) with $a_n = (-1)^n$. □

Exercise 16.7 ([Rud76] 3.2)**.** Calculate $\lim_{n\to\infty}\left(\sqrt{n^2+n}-n\right)$.

Solution. □

Exercise 16.8 ([Rud76] 3.3)**.** The sequence (a_n) is recursively defined by

$$\begin{cases} a_0 = \sqrt{2}, \\ a_{n+1} = \sqrt{2+a_n} \quad n \geq 0. \end{cases}$$

Show that (a_n) converges.

Solution. We first prove by induction that $a_n \leq a_{n+1} \leq 2$ for all $n \in \mathbb{N}$. For $n = 0$,

$$a_0 = \sqrt{2} \leq \sqrt{2+\sqrt{2}} = a_1 \leq \sqrt{2+\sqrt{4}} = 2.$$

If $a_{n-1} \leq a_n \leq 2$, then

$$a_n = \sqrt{2+a_{n-1}} \leq \sqrt{2+a_n} = a_{n+1} \leq \sqrt{2+2} = 2.$$

Hence (a_n) is monotonically increasing and bounded above by 2. By the monotone convergence theorem, (a_n) converges; let $a_n \to a$. Taking the limit on both sides of $a_{n+1} = \sqrt{2+a_n}$,

$$\begin{aligned} \lim_{n\to\infty} a_{n+1} &= \lim_{n\to\infty} \sqrt{2+a_n} \\ a &= \sqrt{2+a} \\ a &= 2 \text{ or } 1 \end{aligned}$$

Since all $a_n \geq 0$, we must have $a = 2$. □

Exercise 16.9 (Contractive sequence)**.** A complex sequence (x_n) is *contractive* if there exists $k \in [0,1)$ such that

$$|a_{n+2} - a_{n+1}| \leq k|a_{n+1} - a_n| \quad (\forall n \in \mathbb{N})$$

Show that every contractive sequence is convergent.

Solution. By induction on n, we have

$$|a_{n+1} - a_n| \leq k^{n-1}|a_2 - a_1| \quad (\forall n \in \mathbb{N})$$

Thus

$$\begin{aligned}|a_{n+p}-a_n| &\le |a_{n+1}-a_n|+|a_{n+2}-a_{n+1}|+\cdots+|a_{n+p}-a_{n+p-1}|\\ &\le \left(k^{n-1}+k^n+\cdots+k^{n+p-2}\right)|a_2-a_1|\\ &\le k^{n-1}\left(1+k+k^2+\cdots+k^{p-1}\right)|a_2-a_1|\\ &\le \frac{k^{n-1}}{1-k}|a_2-a_1|\end{aligned}$$

for all $n,p\in\mathbb{N}$. Since $k^{n-1}\to 0$ as $n\to\infty$ (independently of p), this implies (a_n) is a Cauchy sequence, so it is convergent. □

Exercise 16.10 ([Rud76] 3.4)**.** Find the limit superior and limit inferior of the sequence (a_n) defined by

$$a_1=0,\quad a_{2m}=\frac{a_{2m-1}}{2},\quad a_{2m+1}=a_{2m}+\frac{1}{2}.$$

Solution. We shall prove by induction that

$$a_{2m}=\frac{1}{2}-\frac{1}{2^m},\quad a_{2m+1}=1-\frac{1}{2^m}$$

for $m=1,2,\ldots$. The second of these equalities is a direct consequence of the first, and so we need only prove the first. Immediate computation shows that $a_2=0$ and $a_3=\frac{1}{2}$. Hence assume that both formulae holds for $m\le r$. Then

$$a_{2r+2}=\frac{1}{2}a_{2r+1}=\frac{1}{2}\left(1-\frac{1}{2^r}\right)=\frac{1}{2}-\frac{1}{2^{r+1}}.$$

This completes the induction. We thus have $\limsup_{n\to\infty} a_n=1$ and $\liminf_{n\to\infty} a_n=\frac{1}{2}$. □

Exercise 16.11 ([Rud76] 3.7)**.** Prove that the convergence of $\sum a_n$ implies the convergence of

$$\sum\frac{\sqrt{a_n}}{n}$$

if $a_n\ge 0$.

Exercise 16.12 ([Rud76] 3.13)**.** Prove that the Cauchy product of two absolutely convergent series converges absolutely.

Exercise 16.13 ([Rud76] 3.23)**.** Suppose (a_n) and (b_n) are Cauchy sequences in a metric space X. Show that the sequence $(d(a_n,b_n))$ converges.

Exercise 16.14. Prove that $\sum_{n=1}^{\infty} \frac{\sin n\alpha}{n}$ is convergent for all real α, and that $\sum_{n=1}^{\infty} \frac{\cos n\alpha}{n}$ is convergent if and only if $\alpha \neq 2k\pi$ for integer k.
Hint:

$$\begin{aligned}
&\sin x + \sin 2x + \sin 3x \\
&= \frac{\left(\sin \frac{x}{2}\right)\left(\sin x + \sin 2x + \sin 3x\right)}{\sin \frac{x}{2}} \\
&= \frac{-\frac{1}{2}\left(\cos \frac{3x}{2} - \cos \frac{x}{2} - \cdots\right)}{\sin \frac{x}{2}}
\end{aligned}$$

Exercise 16.15. Determine whether the series

$$\sum_{n=1}^{\infty} \ln\left(1 + \frac{(-1)^n}{n^p}\right)$$

is absolutely convergent, conditionally convergent or divergent for each $p > 0$.
Hint: $\ln(1+x) \to x$ as $x \to 0$.

Exercise 16.16. Suppose that $\sum_{n=1}^{\infty} a_n$ is absolutely convergent; for $n = 1$ to ∞, define the subsequences

$$b_n = a_{2n-1}, \quad c_n = a_{2n}.$$

(i) Show that $\sum b_n$ and $\sum c_n$ are absolutely convergent.

(ii) Show that $\sum a_n = \sum b_n + \sum c_n$.

(iii) Using the fact that $\sum \frac{1}{n^2} = \frac{\pi^2}{6}$, find $\sum \frac{(-1)^{n-1}}{n^2}$.

Exercise 16.17. Given a function $f\colon \mathbb{R} \to \mathbb{R}$ such that for any $x \in \mathbb{R}$, there exists $\varepsilon > 0$ such that f is monotically increasing on $(x-\varepsilon, x+\varepsilon)$. Prove that f is monotonically increasing on $\mathbb{R}$.

17 Continuity

Let (X, d_X) and (Y, d_Y) be metric spaces. Let $E \subset X$, then the metric d_X induces a metric on E. Consider a function $f\colon E \to Y$. In particular, if $Y = \mathbb{R}$, f is called a *real-valued function*; if $Y = \mathbb{C}$, f is called a *complex-valued function*.

17.1 Limit of Functions

Definitions and Properties

Recall that we have previously defined limits for sequences. Now, we will define limits for functions.

Definition 17.1 (Limit of function)**.** Let p be a limit point of E. We say $\lim_{x \to p} f(x) = q$ if there exists $q \in Y$ such that

$$\forall \varepsilon > 0, \quad \exists \delta > 0, \quad \forall x \in E, \quad 0 < d_X(x, p) < \delta \implies d_Y(f(x), q) < \varepsilon. \tag{17.1}$$

Let us examine what (17.1) means. We can get $f(x)$ as close to q as desired, by choosing x sufficiently close to, but not equal to, p.

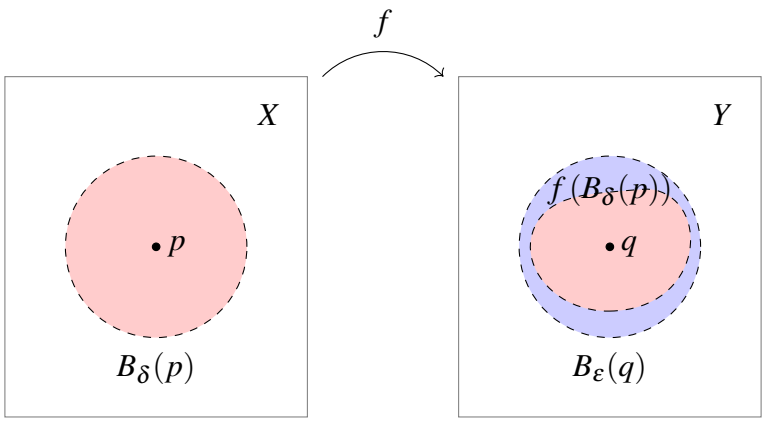


Figure 17.1: Limit of function

Remark. Note that $p \in X$, but it is not necessary that $p \in E$ in the above definition. Moreover, even if $p \in E$, we may very well have $f(p) \neq \lim_{x \to p} f(x)$.

Example (Constant function)**.** Define $f : X \to Y$ by $f(x) = c$ for some constant $c \in Y$. Then for any $p \in X$,

$$\lim_{x \to p} f(x) = c.$$

Proof. Let $\varepsilon > 0$ be given. Then for all $x \in X$,

$$d(f(x), c) = d(c, c) = 0 < \varepsilon,$$

so we can pick δ to be any positive number, such that

$$0 < d(x, p) < \delta \implies d(f(x), c) < \varepsilon.$$

□

Example (Identity function)**.** Define $f : X \to Y$ by $f(x) = x$ for all $x \in X$. Then for any $p \in X$,

$$\lim_{x \to p} f(x) = p.$$

Proof. Let $\varepsilon > 0$ be given. Choose $\delta = \varepsilon$. Then

$$0 < d(x, p) < \delta \implies d(f(x), p) = d(x, p) < \delta = \varepsilon.$$

□

We can recast Definition 17.1 in terms of limits of sequences:

Lemma 17.2 (Sequential criterion for limits)**.** *Let p be a limit point of E. Then*

$$\lim_{x \to p} f(x) = q \tag{I}$$

if and only if

$$\lim_{n \to \infty} f(p_n) = q \tag{II}$$

for every sequence (p_n) in $E \setminus \{p\}$ where $p_n \to p$.

Proof.

$\boxed{\implies}$ Suppose (I) holds. Then fix $\varepsilon > 0$, there exists $\delta > 0$ such that for all $x \in E$,

$$0 < d_X(x, p) < \delta \implies d_Y(f(x), q) < \varepsilon.$$

Let (p_n) be a sequence in $E \setminus \{p\}$. Since $p_n \to p$, for the same $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $n \geq N$,

$$0 < d_X(p_n, p) < \delta.$$

This implies that for $n \geq N$, $d_Y(f(p_n), q) < \varepsilon$. Hence by definition $\lim_{n\to\infty} f(p_n) = q$.

$\boxed{\Longleftarrow}$ Suppose, for a contradiction, (II) holds and (I) does not hold. Then $\lim_{x\to p} f(x) \neq q$, so

$$\exists \varepsilon > 0, \quad \forall \delta > 0, \quad \exists x \in E, \quad 0 < d_X(x, p) < \delta \quad \text{and} \quad d_Y(f(x), q) \geq \varepsilon.$$

Since (II) holds, taking $\delta_n = \frac{1}{n}$ $(n = 1, 2, \ldots)$, we thus find a sequence (p_n) in $E \setminus \{p\}$ such that

$$0 < d_X(p_n, p) < \frac{1}{n} \quad \text{and} \quad d_Y(f(p_n), q) \geq \varepsilon.$$

Clearly $p_n \to p$ but $f(p_n) \not\to q$, contradicting (II). □

Corollary 17.3. *If f has a limit at p, this limit is unique.*

Proof. Suppose $\lim_{x\to p} f(x) = q$ and $\lim_{x\to p} f(x) = q'$. We will show that $q = q'$.

By 17.2, for every sequence (p_n) in $E \setminus \{p\}$ where $p_n \to p$, we have

$$f(p_n) \to q \quad \text{and} \quad f(p_n) \to q'.$$

But the limit of a sequence is unique, so we must have $q = q'$. □

Suppose $f, g \colon E \to \mathbb{C}$. Define

$$(f+g)(x) = f(x) + g(x) \quad (x \in E).$$

We define the difference $f - g$, the product fg, and the quotient f/g similarly, with the understanding that the quotient is defined only at $x \in E$ at which $g(x) \neq 0$.

Similarly, if $\mathbf{f}, \mathbf{g} \colon E \to \mathbb{R}^k$, we define

$$(\mathbf{f}+\mathbf{g})(x) = \mathbf{f}(x) + \mathbf{g}(x), \quad (\mathbf{f} \cdot \mathbf{g})(x) = \mathbf{f}(x) \cdot \mathbf{g}(x);$$

and if λ is a real number, $(\lambda \mathbf{f})(x) = \lambda \mathbf{f}(x)$.

Lemma 17.4 (Limit laws). *Suppose $E \subset X$, p is a limit point of E. Let $f, g \colon E \to \mathbb{C}$, $\lim_{x\to p} f(x) = A$, $\lim_{x\to p} g(x) = B$. Then*

(i) $\lim_{x\to p} (f+g)(x) = A + B$ *(sum)*

(ii) $\lim_{x\to p} (fg)(x) = AB$ *(product)*

(iii) $\lim_{x\to p}\left(\frac{f}{g}\right)(x) = \frac{A}{B}\ (B \neq 0)$ *(quotient)*

Proof. These follow from 17.2 and analagous limit properties of sequences in $\mathbb{C}$. □

If $\mathbf{f}, \mathbf{g}\colon E \to \mathbb{R}^k$, then (i) remains true, and (ii) becomes $\lim_{x\to p}(\mathbf{f}\cdot\mathbf{g})(x) = \mathbf{A}\cdot\mathbf{B}$.

Example. To find

$$\lim_{x\to 2}\frac{2x^3+5x^2-8x-20}{x^3-8}$$

we may write this function as

$$\frac{(x-2)(x+2)(2x+5)}{(x-2)(x^2+2x+4)}$$

and note that if the factor $(x-2)$ is removed, the result is continuous at $x=2$, arriving at

$$\lim_{x\to 2}\frac{(x+2)(2x+5)}{x^2+2x+4} = \frac{(4)(9)}{12} = 3.$$

Infinite Limits and Limits at Infinity

To enable us to operate in the extended real number system, we shall now enlarge the scope of Definition 17.1.

Definition 17.5 (Infinite limit)**.** We write $\lim_{x\to p} f(x) = \infty$ if

$$\forall M \in \mathbb{R}^+, \quad \exists \delta > 0, \quad \forall x \in E, \quad 0 < |x-p| < \delta \implies f(x) > M.$$

Example.

$$\lim_{x\to 1}\frac{1}{(x-1)^2} = \infty$$

$$\lim_{x\to 0}\frac{1}{\sin x^2} = \infty.$$

Definition 17.6 (Limit at infinity)**.** Suppose $f\colon E \subset \mathbb{R} \to \mathbb{R}$, where E is an unbounded interval such as $0 < x < \infty$. We write $\lim_{x\to\infty} f(x) = q$, if

$$\forall \varepsilon > 0, \quad \exists x_0, \quad \forall x \in E, \quad x > x_0 \implies |f(x) - q| < \varepsilon.$$

The analogue of Theorem 4.4 is still true, and the proof offers nothing new. We state it, for the sake of completeness.

Lemma 17.7. *Let $f, g: E \subset \mathbb{R} \to \mathbb{R}$. Suppose $\lim_{t\to x} f(t) = A$, $\lim_{t\to x} g(t) = B$. Then*

(i) $\lim_{t\to x}(f+g)(t) = A+B$

(ii) $\lim_{t\to x}(fg)(t) = AB$

(iii) $\lim_{t\to x}(f/g)(t) = A/B$

provided the RHS are defined.

Note that $\infty - \infty$, $0\cdot\infty$, ∞/∞, $A/0$ are not defined (see Definition 1.23).

One additional refinement of considerable usefulness is the notion of a one-sided limit.

Definition 17.8 (Left-hand limit). Let $f: (a,b) \to \mathbb{R}$. Let $p \in (a,b]$. We write $\lim_{x\to p^-} f(x) = q$, if

$$\forall \varepsilon > 0, \quad \exists \delta > 0, \quad \forall x \in E, \quad p - \delta < x < p \implies |f(x) - q| < \varepsilon.$$

Similarly, let $p \in [a,b)$. We write $\lim_{x\to p^+} f(x) = q$, if

$$\forall \varepsilon > 0, \quad \exists \delta > 0, \quad \forall x \in E, \quad p < x < p + \delta \implies |f(x) - q| < \varepsilon.$$

Notation. Sometimes we write $f(x+)$ in place of $\lim_{t\to x^+} f(t)$, and $f(x-)$ in place of $\lim_{t\to x^-} f(t)$.

Compare this definition with Definition 17.1; for one-sided limits, we are only concerned with half open balls around p (since we only require x to approach p from either the left or right side).

Remark. An equivalent formulation using limits of sequences is presented in [Rud76].

Both may exist when the usual two-sided limit does not, and the two-sided limit exists when and only when both left- and right-hand limits exist and are the same.

Example. The function $f(x) = \dfrac{1}{1+e^{1/x}}$ has left-hand limit 1 and right-hand limit 0 at $x = 0$.

The one-sided limit notation can be combined with the other conventions already introduced.

Example. The type of behaviour exhibited by $f(x) = \frac{1}{x}$ at $x = 0$ can now be described by writing

$$\lim_{x\to 0^+} f(x) = \infty, \qquad \lim_{x\to 0^-} f(x) = -\infty.$$

17.2 Continuous Functions

Definition 17.9 (Continuity). Suppose $E \subset X$. We say $f\colon E \to Y$ is ***continuous*** at $p \in E$, if

$$\forall \varepsilon > 0, \quad \exists \delta > 0, \quad \forall x \in E, \quad d_X(x,p) < \delta \implies d_Y(f(x), f(p)) < \varepsilon. \tag{17.2}$$

If f is continuous at every point of E, we say f is *continuous on* E.

Let us examine what (17.2) means. For any target distance from $f(p)$, we can always find points $x \in E$ sufficiently close to p such that their images under f are within the target distance from $f(p)$.

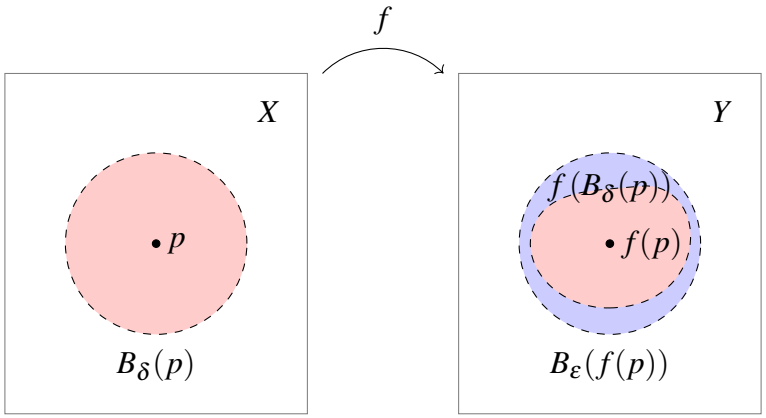


Figure 17.2: Continuity of a function

Remark. For f to be continuous at p, we require f to be defined at p. (Compare this with the remark following Definition 17.1.)

Notation. Let X and Y be metric spaces. The space of continuous bounded functions from X to Y is denoted by $\mathscr{C}(X,Y)$; if $Y = \mathbb{C}$, we simply write $\mathscr{C}(X)$.

Instead of using Definition 17.9, we will often use the next result to show that a function is continuous at a point.

Lemma 17.10. *Let p be a limit point of E. Then f is continuous at p if and only if*

$$\lim_{x \to p} f(x) = f(p).$$

Proof. Compare Definitions 17.1 and 17.9. □

Corollary 17.11 (Sequential criterion for continuity). *$f\colon E \subset X \to Y$ is continuous on*

E if and only if for every convergent sequence (p_n) in E,

$$\lim_{n\to\infty} f(p_n) = f\left(\lim_{n\to\infty} p_n\right).$$

Remark. This means that for continuous functions, the limit symbol can be interchanged with the function symbol. Some care is needed in interchanging these symbols because sometimes $(f(p_n))$ converges when (p_n) diverges.

Lemma 17.12. *Let $f, g: X \to \mathbb{C}$ be continuous on X. Then the following are continuous on X:*

(i) $f+g$ *(sum)*

(ii) fg *(product)*

(iii) f/g *(provided $g(x) \neq 0$ for all $x \in X$)* *(quotient)*

Proof. At isolated points of X, there is nothing to prove.

At limit points, the statement follows from 17.4 and 17.10. □

Example. It is a trivial exercise to show that the following complex-valued functions are continuous on $\mathbb{C}$:

- constant functions, defined by $f(z) = c$ for all $z \in \mathbb{C}$;
- the identity function, defined by $f(z) = z$ for all $z \in \mathbb{C}$.

Repeated application of the previous result establishes the continuity of every polynomial

$$f(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_n z^n$$

where $a_i \in \mathbb{C}$.

We now prove the analogue for Euclidean spaces.

Lemma 17.13.

(i) Let $f_1, \ldots, f_k: X \to \mathbb{R}$, and let $\mathbf{f}: X \to \mathbb{R}^k$ be defined by

$$\mathbf{f}(x) = (f_1(x), \ldots, f_k(x)) \quad (x \in X).$$

Then $\mathbf{f}$ is continuous if and only if each of its components $f_1, \ldots, f_k$ is continuous.

(ii) Let $\mathbf{f}, \mathbf{g}: X \to \mathbb{R}^k$ be continuous on X. Then $\mathbf{f}+\mathbf{g}$ and $\mathbf{f}\cdot\mathbf{g}$ are continuous on X.

Proof. (i) follows from the inequalities

$$|f_j(x) - f_j(y)| \leq |\mathbf{f}(x) - \mathbf{f}(y)| = \left(\sum_{i=1}^{k} |f_i(x) - f_i(y)|^2 \right)^{1/2}$$

for $j = 1, \ldots, k$.

(ii) follows from (i) and 17.12. □

We now consider the composition of functions. The following result shows that a continuous function of a continuous function is continuous.

Proposition 17.14. *Suppose X, Y, Z are metric spaces, $E \subset X$. Let*

- $f: E \to Y$,
- $g: f(E) \subset Y \to Z$,
- $h: E \to Z$ *is defined by* $h = g \circ f$.

If f is continuous at $p \in E$, and g is continuous at $f(p)$, then h is continuous at p.

Proof. Let $\varepsilon > 0$ be given. Since g is continuous at $f(p)$, there exists $\eta > 0$ such that for all $y \in f(E)$,

$$d_Y(y, f(p)) < \eta \implies d_Z(g(y), g(f(p))) < \varepsilon. \tag{I}$$

Since f is continuous at p, there exists $\delta > 0$ such that for all $x \in E$,

$$d_X(x, p) < \delta \implies d_Y(f(x), f(p)) < \eta. \tag{II}$$

Combining (I) and (II), it follows that for all $x \in E$,

$$d_X(x, p) < \delta \implies d_Z(h(x), h(p)) = d_Z(g(f(x)), g(f(p))) < \varepsilon.$$

Therefore h is continuous at p. □

Notation. While functions are technically defined on a subset E of a metric space, the complement of E plays no role in the definition of continuity, so we can safely ignore the complement, and think of continuous functions as mappings from one metric space to another.

Continuity and Pre-images of Open or Closed Sets

The next result is a useful characterisation of continuity: a function is continuous if and only if the pre-image of every open set is open.

Lemma 17.15. *$f: X \to Y$ is continuous on X if and only if $f^{-1}(U)$ is open in X for every open set $U \subset Y$.*

Proof.

$\boxed{\Longrightarrow}$ Suppose f is continuous on X. Let $U \subset Y$ be open. Let $p \in f^{-1}(U)$.

Since $p \in f^{-1}(U)$, there exists $y \in U$ such that $f(p) = y$. By openness of U, there exists $\varepsilon > 0$ such that $B_\varepsilon(y) \subset U$.

Since f is continuous at p, for the same ε, there exists $\delta > 0$ such that for all $x \in X$,

$$d_X(x,p) < \delta \implies d_Y(f(x),y) < \varepsilon,$$

or

$$f(B_\delta(p)) \subset B_\varepsilon(y).$$

Hence

$$B_\delta(p) \subset f^{-1}(f(B_\delta(p))) \subset f^{-1}(B_\varepsilon(y)) \subset f^{-1}(U),$$

so $f^{-1}(U)$ is open in X.

$\boxed{\Longleftarrow}$ Suppose $f^{-1}(U)$ is open in X for every open set $U \subset Y$. Fix $p \in X$, let $y = f(p)$. We will show that f is continuous at p.

For every $\varepsilon > 0$, the ball $B_\varepsilon(y)$ is open in Y, so $f^{-1}(B_\varepsilon(y))$ is open in X (by assumption). Now $p \in f^{-1}(B_\varepsilon(y))$, so by openness of $f^{-1}(B_\varepsilon(y))$, there exists $\delta > 0$ such that $B_\delta(p) \subset f^{-1}(B_\varepsilon(y))$. Hence $f(B_\delta(p)) \subset B_\varepsilon(y)$; that is,

$$d_X(x,p) < \delta \implies d_Y(f(x),y) < \varepsilon.$$

Therefore f is continuous at p. □

Similarly, a function is continuous if and only if the pre-image of every closed set is closed.

Corollary 17.16. *$f: X \to Y$ is continuous on X if and only if $f^{-1}(C)$ is closed in X for every closed set $C \subset Y$.*

Proof. This follows from the above result, since a set is closed if and only if its complement is open, and since $f^{-1}(E^c) = [f^{-1}(E)]^c$ for every $E \subset Y$. □

Continuity and Compactness

We say $\mathbf{f}: E \to \mathbb{R}^k$ is *bounded* if there exists $M \in \mathbb{R}$ such that $\|\mathbf{f}(x)\| \leq M$ for all $x \in E$.

The next result shows that continuous functions preserve compactness.

Proposition 17.17. *Suppose $f\colon X \to Y$ is continuous on X, where X is compact. Then $f(X)$ is compact.*

Proof. Let $\mathscr{U} = \{U_i\}_{i\in I}$ be an open cover of $f(X)$. Since f is continuous on X, by 17.15, each of the sets $f^{-1}(U_i)$ is open.

Consider the open cover $\{f^{-1}(U_i)\}_{i\in I}$. Since X is compact, there exist finitely many indices $i_1,\dots,i_n$ such that

$$X \subset \bigcup_{k=1}^{n} f^{-1}(U_{i_k}).$$

Since $f(f^{-1}(E)) \subset E$ for every $E \subset Y$,

$$f(X) \subset f\left(\bigcup_{k=1}^{n} f^{-1}(U_{i_k})\right) = f\left(f^{-1}\left(\bigcup_{k=1}^{n} U_{i_k}\right)\right) \subset \bigcup_{k=1}^{n} U_{i_k}.$$

Hence the sets $\{U_{i_1},\dots,U_{i_n}\}$ form a finite subcover of $\mathscr{U}$, so $f(X)$ is compact. □

Corollary 17.18. *If $\mathbf{f}\colon X \to \mathbb{R}^k$ is continuous on X, where X is compact, then $\mathbf{f}(X)$ is closed and bounded. Thus, $\mathbf{f}$ is bounded.*

Proof. By 17.17, $\mathbf{f}(X)$ is compact. Since $\mathbf{f}(X) \subset \mathbb{R}^k$, by the Heine–Borel theorem, $\mathbf{f}(X)$ is closed and bounded. □

The result is particularly important when f is a real-valued function; the next result states that a continuous real-valued function on a compact set must attain its minimum and maximum.

Corollary 17.19 (Extreme value theorem)**.** *Suppose $f\colon X \to \mathbb{R}$ is continuous, X is compact. Let*

$$M = \sup_{p\in X} f(p), \quad m = \inf_{p\in X} f(p).$$

Then there exist $p,q \in X$ such that $f(p) = M$ and $f(q) = m$.

Proof. From the previous corollary, $f(X)$ is a closed and bounded set in $\mathbb{R}$. Hence $f(X)$ contains its supremum and infimum, by 15.25. □

Proposition 17.20. *Suppose $f\colon X \to Y$ is continuous on X and bijective, X is compact. Then its inverse $f^{-1} : Y \to X$ is continuous on Y.*

Proof. By 17.15, it suffices to prove that $f(U)$ is open in Y for every open set U in X.

Let U be an open set in X. Then its complement U^c is closed in X. Since U^c is a closed subset of a compact set X, U^c is compact. Thus by 17.17, $f(U^c)$ is a compact subset of Y, so $f(U^c)$ is closed in Y.

Since f is bijective and thus surjective, $f(U)$ is the complement of $f(U^c)$. Hence $f(U)$ is open. □

Bolzano's Theorem

Lemma 17.21 (Sign-preserving property). *Let $f\colon [a,b] \to \mathbb{R}$ be continuous at $c \in [a,b]$, $f(c) \neq 0$. Then there exists $\delta > 0$ such that $f(x)$ has the same sign as $f(c)$ for $c - \delta < x < c + \delta$.*

Proof. Assume $f(c) > 0$. Let $\varepsilon > 0$ be given. By continuity of f, there exists $\delta > 0$ such that

$$c - \delta < x < c + \delta \implies f(c) - \varepsilon < f(x) < f(c) + \varepsilon.$$

Take the δ corresponding to $\varepsilon = \frac{f(c)}{2}$. Then

$$\frac{1}{2} f(c) < f(x) < \frac{3}{2} f(c) \quad (c - \delta < x < c + \delta)$$

so $f(x)$ has the same sign as $f(c)$ for $c - \delta < x < c + \delta$.

The proof is similar if $f(c) < 0$, except that we take $\varepsilon = -\frac{1}{2} f(c)$. □

The next result states that if the graph of $f\colon [a,b] \to \mathbb{R}$ lies above the x-axis at a and below the x-axis at b, then the graph must cross the axis somewhere in between. (This should be intuitively obvious.)

Theorem 17.22 (Bolzano). *Suppose $f\colon [a,b] \to \mathbb{R}$ is continuous, and $f(a)f(b) < 0$ (that is, $f(a)$ and $f(b)$ have opposite signs). Then there exists $c \in (a,b)$ such that $f(c) = 0$.*

Proof. For definiteness, assume $f(a) > 0$ and $f(b) < 0$. Let

$$A = \{x \in [a,b] \mid f(x) \geq 0\}.$$

Then A is non-empty since $a \in A$, and A is bounded above by b, so A has a supremum in $\mathbb{R}$; let $c = \sup A$. Then $a < c < b$.

Claim. $f(c) = 0$.

If $f(c) \neq 0$, by the previous result, there exists $\delta > 0$ such that $f(x)$ has the same sign as $f(c)$ for $c - \delta < x < c + \delta$.

- If $f(c) > 0$, there are points $x > c$ at which $f(x) > 0$, contradicting the definition of c.
- If $f(c) < 0$, then $c - \frac{\delta}{2}$ is an upper bound for A, again contradicting the definition of c.

Therefore we must have $f(c) = 0$. □

Continuity and Connectedness

The next result states that under a continuous mapping, the image of any connected set is connected.

Proposition 17.23. *Suppose $f: X \to Y$ is continuous. If $E \subset X$ is connected, then $f(E)$ is connected.*

Proof. We prove the contrapositive. Suppose $f(E)$ is not connected, then $f(E) = A \cup B$ for some $A, B \subset Y$ where $\overline{A} \cap B = \overline{B} \cap A = \emptyset$.

Consider $\overline{A}$ and $\overline{B}$, which are closed in Y. Since f is continuous, by 17.16, $f^{-1}(\overline{A})$ and $f^{-1}(\overline{B})$ are closed in X; let $K_A = f^{-1}(\overline{A})$, $K_B = f^{-1}(\overline{B})$. We now want to construct a separation of E.

Let $E_1 = f^{-1}(A) \cap E$, $E_2 = f^{-1}(B) \cap E$. Since $A \cap B = \emptyset$, we have that $E_1 \cap E_2 = \emptyset$. Since $A, B \neq \emptyset$, we have that $E_1, E_2 \neq \emptyset$.

Claim. E_1 and E_2 is a separation of E.

Notice $E_1 \subset K_A$ (which is closed) and $E_2 \subset K_B$ (which is closed). Then $\overline{E_1} \subset K_A$ and $\overline{E_2} \subset K_B$. Note that

$$f^{-1}(\overline{A}) \cap f^{-1}(B) = f^{-1}(\overline{A} \cap B) = \emptyset$$

so $K_A \cap E_2 = \emptyset$. Similarly $K_B \cap E_1 = \emptyset$.

Therefore E is separated. □

An important corollary states that a continuous real-valued function assumes all intermediate values on an interval.

Corollary 17.24 (Intermediate value theorem)**.** *Suppose $f: [a,b] \to \mathbb{R}$ is continuous. If $f(a) < f(b)$ and $f(a) < c < f(b)$, then there exists $x \in (a,b)$ such that $f(x) = c$.*

Proof. By 15.57, $[a,b]$ is connected. By the previous result, $f([a,b])$ is a connected subset of $\mathbb{R}$. Then apply 15.58 and we are done. □

Remark. The converse is not necessarily true. For instance, the *topologist's sine curve*

$$f(x) = \begin{cases} 0 & (x = 0) \\ \sin\left(\frac{1}{x}\right) & (x \neq 0) \end{cases}$$

satisfies the intermediate value property, but f is not continuous.

Example. $x^5 - 2x^3 + 3x^2 - 1 = 0$ has a solution in the interval $[0,1]$.

Proof. Let $f(x) = x^5 - 2x^3 + 3x^2 - 1$. f is continuous on $[0,1]$. In addition,

$$f(0) = -1 < 0, \quad 0 < 1 = f(1).$$

By the intermediate value theorem, there exists $x \in (0,1)$ such that $x^5 - 2x^3 + 3x^2 - 1 = 0$. □

One-sided Continuity

Similar to defining one-sided limits, we can define continuity from one direction.

Definition 17.25. Let $f\colon [a,b] \to \mathbb{R}$.
If f is defined at $x \in (a,b]$ and $f(x-) = f(x)$, we say f is *continuous from the left* at x.
If f is defined at $x \in [a,b)$ and $f(x+) = f(x)$, we say f is *continuous from the right* at x.

Lemma 17.26. *Let $f : (a,b) \to \mathbb{R}$. Then f is continuous at x if and only if*

$$f(x-) = f(x) = f(x+).$$

That is, f is continuous at x if and only if f is continuous from the left and right at x.

Proof.

$\boxed{\Longrightarrow}$

$\boxed{\Longleftarrow}$ Suppose f is left and right continuous at x.

Let $\varepsilon > 0$ be given. Then there exists $\delta_l > 0$ and $\delta_r > 0$ such that

$$x - \delta_l < t \leq x \implies |f(t) - f(x)| < \varepsilon$$
$$x < t \leq x + \delta_r \implies |f(t) - f(x)| < \varepsilon$$

Choose $\delta = \min\{\delta_l, \delta_r\}$. Then if $|t - x| < \delta$, either $t \leq x$ or $t \geq x$. In the first case, $t > x - \delta > x - \delta_l$ and so $|f(t) - f(x)| < \varepsilon$. In the second case, $t < x + \delta < x + \delta_r$ and so $|f(t) - f(x)| < \varepsilon$ and the function is continuous at x. □

17.3 Uniform Continuity

Definition 17.27 (Uniform continuity)**.** We say $f: X \to Y$ is ***uniformly continuous*** on X, if

$$\forall \varepsilon > 0, \quad \exists \delta > 0, \quad \forall p, q \in X, \quad d_X(p,q) < \delta \implies d_Y(f(p), f(q)) < \varepsilon. \tag{17.3}$$

There is a vivid way to describe this property of a function that may help to emphasise its characteristic features. Consider a measuring device with two movable prongs and a meter which indicates the temperature difference of the prongs. Suppose we have a sheet of metal whose temperature varies from point to point, and the temperature at p is $f(p)$. If we place one prong at a point p on the sheet and the other at q, the meter reads the value $d(f(p), f(q))$.

If f is uniformly continuous on X, this means that for a given value $\varepsilon > 0$, we can set the prongs at a fixed separation of at most δ and be sure that the meter will read no higher than ε.

Remark. The difference between continuity and uniform continuity is that of one between a local and global property.

- Continuity can be defined at a single point, as δ depends on ε as well as the point p.
- Uniform continuity is a property of a function on a set, as the same δ has to work for *all* $p \in X$ (which ensures a *uniform* rate of closeness across the entire domain.).

Hence uniform continuity is a stronger continuity condition than continuity; a function that is uniformly continuous is continuous but a function that is continuous is not necessarily uniformly continuous.

Example. $f(x) = \frac{1}{x}$ is not uniformly continuous on $(0, 1]$.

Proof. Let $\varepsilon = 10$, and suppose we could find a δ ($0 \leq \delta < 1$) that satisfies (17.3). Taking $p = \delta$, $q = \frac{\delta}{11}$, we obtain $|p - q| < \delta$ and

$$|f(p) - f(q)| = \frac{11}{\delta} - \frac{1}{\delta} = \frac{10}{\delta} > 10 = \varepsilon.$$

This contradicts the definition of uniform continuity. □

Example. $f(x) = x^2$ is uniformly continuous on $(0, 1]$.

Proof. Let $\varepsilon > 0$ be given. Take $\delta = \frac{\varepsilon}{2}$. Then for every $p, q \in (0, 1]$, $|p - q| < \delta$ implies

$$|f(p) - f(q)| = |p^2 - q^2| = |(p+q)(p-q)| < 2|p - q| = \varepsilon.$$

□

Evidently every uniformly continuous function is continuous. For the converse to hold, we need an additional condition.

Proposition 17.28. *On a compact set, every continuous function is uniformly continuous.*

Proof. Suppose $f\colon X \to Y$ is continuous on X, where X is compact. We will show that f is uniformly continuous on X.

Let $\varepsilon > 0$ be given. Take $p \in X$. Since f is continuous at p, there exists $\delta_p > 0$ such that for all $q \in X$,

$$d_X(p,q) < \delta_p \implies d_Y(f(p),f(q)) < \frac{\varepsilon}{2}.$$

Consider the collection of open balls centred at each $p \in X$:

$$\left\{ B_{\frac{1}{2}\delta_p}(p) \,\middle|\, p \in X \right\}.$$

This forms an open cover of X. Since X is compact, there exists finitely many points $p_1, \ldots, p_n \in X$ such that

$$X \subset \bigcup_{i=1}^{n} B_{\frac{1}{2}\delta_{p_i}}(p_i).$$

Claim. Take $\delta = \min\left\{ \frac{1}{2}\delta_{p_1}, \ldots, \frac{1}{2}\delta_{p_n} \right\}$.

Let $p,q \in X$ be such that $d_X(p,q) < \delta$. Since X is covered by finitely many open balls, we have $p \in B_{\frac{1}{2}\delta_{p_i}}(p_i)$ for some $i \in \{1,\ldots,n\}$. Thus

$$d_X(p,p_i) < \frac{1}{2}\delta_{p_i}.$$

We also have

$$d_X(q,p_i) \leq d_X(p,q) + d_X(p,p_i) < \delta + \frac{1}{2}\delta_{p_i} \leq \delta_{p_i}.$$

By the continuity of f,

$$\begin{aligned} d_Y(f(p),f(q)) &\leq d_Y(f(p),f(p_m)) + d_Y(f(q),f(p_m)) \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

Hence f is uniformly continuous on X. □

Lemma 17.29 (Lebesgue covering lemma). *Suppose $\{U_i \mid i \in I\}$ is an open cover of a compact metric space X. Then there exists $\delta > 0$ such that for all $x \in X$,*

$$B_\delta(x) \subset U_i$$

for some $i \in I$; we say δ is a Lebesgue number *of the cover.*

Proof. Since X is compact, there exist finitely many indices $i_1, \ldots, i_n$ such that

$$X \subset \bigcup_{k=1}^{n} U_{i_k}.$$

If any one of the U_{i_k}'s equals X, then any $\delta > 0$ will serve as a Lebesgue number.

For any closed set A, define the distance

$$d(x,A) := \inf_{a \in A} d(x,a).$$

Claim. $d(x,A)$ is a continuous function of x.

Proof. Fix a closed set A. Let $x \in X$. Let $\varepsilon > 0$ be given. For all $y \in X$, we want to show that there exists $\delta > 0$ such that $d_X(x,y) < \delta$ implies $|d(x,A) - d(y,A)| < \varepsilon$. □

Consider the average distance from each x to the complements of U_{i_k} as the function of x:

$$f(x) = \frac{1}{n} \sum_{k=1}^{n} d(x, {U_{i_k}}^c).$$

Since f is a sum of continuous functions, f is continuous. Since f is continuous on a compact set, f attains its minimum value; call it δ. See that $\delta > 0$ since $\{U_{i_1}, \ldots, U_{i_n}\}$ is an open cover (so $x \in U_{i_k}$ implies $d(x, {U_{i_k}}^c) > 0$).

For each x, $f(x) \geq \delta$ implies that at least one of the distances $d(x, {U_{i_k}}^c) \geq \delta$. Hence $B_\delta(x) \subset U_{i_k}$, as desired. □

17.4 Discontinuities

If f is not continuous at x, we say that f is *discontinuous* at x, or that f has a *discontinuity* at x.

Example (Dirichlet function)**.** The *Dirichlet function*, defined by

$$f(x) = \begin{cases} 1 & (x \in \mathbb{Q}) \\ 0 & (x \in \mathbb{R} \setminus \mathbb{Q}) \end{cases}$$

is discontinuous everywhere; that is, f is not continuous at any point in $\mathbb{R}$.

Proof. Let $x \in \mathbb{R}$. We consider two cases.

Case 1: $x \in \mathbb{Q}$**.** Then $f(x) = 1$. Take $\varepsilon = \frac{1}{2}$. Since the irrational numbers are dense in the reals, for any $\delta > 0$, we can always find an irrational $y \in \mathbb{R} \setminus \mathbb{Q}$ such that

$$|x - y| < \delta \quad \text{and} \quad |f(x) - f(y)| = 1 \geq \frac{1}{2}.$$

Case 2: $x \in \mathbb{R} \setminus \mathbb{Q}$**.** Then $f(x) = 0$. Again take $\varepsilon = \frac{1}{2}$. Since $\mathbb{Q}$ is dense in $\mathbb{R}$, for any $\delta > 0$, we can always find $y \in \mathbb{Q}$ such that

$$|x - y| < \delta \quad \text{and} \quad |f(x) - f(y)| = 1 \geq \frac{1}{2}.$$

□

If f is defined on an interval, it is customary to divide discontinuities into two types.

Definition 17.30 (Discontinuities)**.** Let $f\colon (a,b) \to \mathbb{R}$. Suppose f is discontinuous at $x \in (a,b)$.

(i) We say f has a *discontinuity of the first kind* (or a *simple discontinuity*) at x, if $f(x+)$ and $f(x-)$ exist;

(ii) we say f has a *discontinuity of the second kind* if otherwise.

There are two ways in which a function can have a simple discontinuity: either $f(x+) \neq f(x)$ [in which case the value $f(x)$ is immaterial], or $f(x+) = f(x-) \neq f(x)$.

Example.

- Define $f\colon (-3,1) \to \mathbb{R}$ by

$$f(x) = \begin{cases} x+2 & (-3 < x < -2) \\ -x-2 & (-2 \leq x < 0) \\ x+2 & (0 \leq x < 1) \end{cases}$$

  has a simple discontinuity at $x = 0$, and is continuous at every other point of $(-3,1)$.

- The Dirichlet function has a discontinuity of the second kind at every $x \in \mathbb{R}$, since both $f(x+)$ and $f(x-)$ do not exist.

- The topologist's sine curve has a discontinuity of the second kind at $x = 0$, since $f(x+)$ does not exist.

17.5 Monotonic Functions

We now study those functions which never decrease (or never increase) on a given interval.

Definition 17.31 (Monotonicity). We say $f\colon (a,b) \to \mathbb{R}$ is

(i) *monotonically increasing*, if $f(x_1) \leq f(x_2)$ for any $a < x_1 \leq x_2 < b$;

(ii) *monotonically decreasing*, if $f(x_1) \geq f(x_2)$ for any any $a < x_1 \leq x_2 < b$;

(iii) ***monotonic*** if it is either monotonically increasing or monotonically decreasing.

Proposition 17.32. *Let $f\colon (a,b) \to \mathbb{R}$ be monotonically increasing. Then $f(x+)$ and $f(x-)$ exist for all $x \in (a,b)$; more precisely,*

$$\sup_{t\in(a,x)} f(t) = f(x-) \leq f(x) \leq f(x+) = \inf_{t\in(x,b)} f(t).$$

Furthermore, if $a < x < y < b$, then

$$f(x+) \leq f(y-).$$

Analogous results evidently hold for monotonically decreasing functions.

Proof. We will prove the first half of the given statement; the second half can be proven in precisely the same way.

Let $x \in (a,b)$. Since f is monotonically increasing, the set

$$A := \{f(t) \mid a < t < x\}$$

is bounded above by the number $f(x)$. Hence A has a supremum in $\mathbb{R}$; let $\alpha = \sup A$. Evidently $\alpha \leq f(x)$.

Claim. $f(x-) = \alpha$.

To prove this, we need to show that for all $\varepsilon > 0$, there exists $\delta > 0$ such that

$$x - \delta < t < x \implies |f(t) - \alpha| < \varepsilon.$$

Let $\varepsilon > 0$ be given. Since $\alpha = \sup A$, there exists $\delta > 0$ such that $a < x - \delta < x$ and

$$\alpha - \varepsilon < f(x - \delta) \leq \alpha. \tag{1}$$

Since f is monotonic, we have

$$f(x-\delta) \leq f(t) \leq \alpha \quad (x-\delta < t < x) \tag{2}$$

Combining (1) and (2) gives

$$|f(t)-\alpha| < \varepsilon \quad (x-\delta < t < x)$$

as desired. Hence $f(x-) = \alpha$.

Next, if $a < x < y < b$, we see from the given statement that

$$f(x+) = \inf_{t\in(x,b)} f(t) = \inf_{t\in(x,y)} f(t)$$

where the last equality is obtained by applying the given statement to (a,y) in place of (a,b). Similarly,

$$f(y-) = \sup_{t\in(a,y)} f(t) = \sup_{t\in(x,y)} f(t).$$

Comparing these two equations, we conclude that $f(x+) \leq f(y-)$. □

Corollary 17.33. *Monotonic functions have no discontinuities of the second kind.*

Proposition 17.34. *Let $f\colon (a,b) \to \mathbb{R}$ be monotonic. Then the set of points of (a,b) at which f is discontinuous is at most countable.*

Proof. Suppose, for the sake of definiteness, that f is monotonically increasing. Let D be the set of points at which f is discontinuous.

For every $x \in D$, we associate a rational number $r(x)$, where

$$f(x-) < r(x) < f(x+).$$

We now check that the rationals picked for two distinct points of discontinuities are different: since $x_1 < x_2$ implies $f(x_1+) \leq f(x_2-)$ (from the previous result), we see that $r(x_1) \neq r(x_2)$ if $x_1 \neq x_2$.

We have thus established a 1-1 correspondence between D and a subset of $\mathbb{Q}$ (which we know is at most countable). Hence D is at most countable. □

17.6 Lipschitz Continuity

Definition 17.35 (Lipschitz continuity)**.** We say $f: X \to Y$ is ***Lipschitz continuous*** if there exists $K \geq 0$ such that

$$\forall x, y \in X, \quad d_Y(f(x), f(y)) \leq K d_X(x, y).$$

K is called a *Lipschitz constant* for f; we also refer to f as *K-Lipschitz*.

Lemma 17.36. *Lipschitz continuity implies uniform continuity.*

Proof. Let $f: X \to Y$ be K-Lipschitz continuous.

Let $\varepsilon > 0$ be given, let $x, y \in X$. We consider two cases.

Case 1: $K \leq 0$**.** Then

$$d_X(x, y) \leq 0 d_Y(f(x), f(y))$$

so

$$d_X(x, y) \leq 0 \implies d_X(x, y) = 0 \implies x = y$$

for all $x, y \in X$. Hence f is a constant function, which is uniformly continuous.

Case 2: $K > 0$**.** Take $\delta = \frac{\varepsilon}{K}$. If $d_X(x, y) < \delta$, then

$$K d_X(x, y) < \varepsilon.$$

By Lipschitz continuity of f,

$$d_Y(f(x), f(y)) \leq K d_X(x, y).$$

These last two statements together imply $d_Y(f(x), f(y)) < \varepsilon$. Hence f is uniformly continuous on X.

□

We say $f: X \to Y$ is a *contraction* if it is a K-Lipschitz map for some $K < 1$.

Let $f: X \to X$, we say $x \in X$ is a *fixed point* if $f(x) = x$.

Theorem 17.37 (Contraction mapping theorem)**.** *Let X be a complete metric space, and $f: X \to X$ be a contraction. Then f has a unique fixed point.*

Remark. The hypotheses "complete" and "contraction" are necessary. For example, $f: (0,1) \to (0,1)$ defined by $f(x) = Kx$ for any $0 < K < 1$ is a contraction with no fixed point. Also, $f: \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x + 1$ is not a contraction ($K = 1$) and has no fixed point.

Proof. Pick any $x_0 \in X$. Define a sequence (x_n) by $x_{n+1} = f(x_n)$. Since f is a contraction, we have

$$\begin{aligned} d(x_{n+1}, x_n) &= d\left(f(x_n), f(x_{n-1})\right) \\ &\leq K d(x_n, x_{n-1}) \\ &\leq \cdots \\ &\leq K^n d(x_1, x_0) \end{aligned}$$

by induction. Suppose $m \geq n$, then

$$\begin{aligned} d(x_m, x_n) &\leq \sum_{i=n}^{m-1} d(x_{i+1}, x_i) \\ &\leq \sum_{i=n}^{m-1} K^i d(x_1, x_0) \\ &= K^n d(x_1, x_0) \sum_{i=0}^{m-n-1} k^i \\ &\leq K^n d(x_1, x_0) \sum_{i=0}^{\infty} K^i = \frac{K^n}{1-K} d(x_1, x_0). \end{aligned}$$

Thus (x_n) is a Cauchy sequence. Since X is complete, (x_n) converges; let $\lim\limits_{n\to\infty} x_n = x$ for some $x \in X$.

Claim. x is our unique fixed point.

Note that f is continuous because it is a contraction. Hence

$$f(x) = \lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} x_{n+1} = x,$$

so x is a fixed point.

Let x' also be a fixed point. Then

$$d(x, x') = d\left(f(x), f(x')\right) = K d(x, x').$$

As $K < 1$ this means that $d(x, x') = 0$ and hence $x = x'$. The theorem is proved. □

Note that the proof is constructive. Not only do we know that a unique fixed point exists. We also know how to find it.

Exercises

Exercise 17.1 ([Rud76] 4.1)**.** Suppose $f\colon \mathbb{R} \to \mathbb{R}$ satisfies

$$\lim_{h\to 0} (f(x+h) - f(x-h)) = 0$$

for every $x \in \mathbb{R}$. Does this imply that f is continuous?

Solution. No. Consider

$$f(x) = \begin{cases} 1 & (x = 0) \\ 0 & (x \neq 0) \end{cases}$$

□

Exercise 17.2 ([Rud76] 4.2)**.** If $f\colon X \to Y$ is continuous, prove that

$$f(\overline{E}) \subset \overline{f(E)}$$

for every $E \subset X$.

Exercise 17.3 ([Rud76] 4.3)**.** Let $f\colon X \to \mathbb{R}$ be continuous. Let the *zero set* of f be

$$Z(f) = \{x \in X \mid f(x) = 0\}.$$

Prove that $Z(f)$ is closed.

Exercise 17.4 ([Rud76] 4.8)**.** Let f be a real uniformly continuous function on the bounded set $E \subset \mathbb{R}$. Prove that f is bounded on E.
Show that the conclusion is false if boundedness of E is omitted from the hypothesis.

Exercise 17.5 ([Rud76] 4.11)**.** Suppose $f\colon X \to Y$ is uniformly continuous on X. Prove that $(f(x_n))$ is a Cauchy sequence in Y for every Cauchy sequence (x_n) in X.

Exercise 17.6 ([Rud76] 4.12)**.** A uniformly continuous function of a uniformly continuous function is uniformly continuous.

Exercise 17.7 ([Rud76] 4.14)**.** Let $I = [0,1]$ be the closed unit interval. Suppose f is a continuous mapping of I into I. Prove that $f(x) = x$ for at least one $x \in I$.

Exercise 17.8 ([Rud76] 4.15)**.** $f: X \to Y$ is said to be *open* if $f(V)$ is an open set in Y whenever V is an open set in X.
Prove that every continuous open mapping of $\mathbb{R}$ into $\mathbb{R}$ is monotonic.

Exercise 17.9 ([Rud76] 4.16)**.** Let $[x]$ denote the largest integer contained in x, and let $\{x\} = x - [x]$ denote the fractional part of x. What discontinuities do the functions $[x]$ and $\{x\}$ have?

Exercise 17.10 ([Rud76] 4.18)**.** Every rational x can be written in the form $x = \frac{m}{n}$, where $m \in \mathbb{Z}$, $n \in \mathbb{N}$, $\gcd(m,n) = 1$. When $x = 0$, we take $n = 1$. Consider the function f defined on $\mathbb{R}$ by

$$f(x) = \begin{cases} 0 & (x \in \mathbb{R} \setminus \mathbb{Q}) \\ \frac{1}{n} & \left(x = \frac{m}{n}\right) \end{cases}$$

Prove that f is continuous at every irrational point, and that f has a simple discontinuity at every rational point.

Exercise 17.11 ([Rud76] 4.26)**.** Suppose X, Y, Z are metric spaces, and Y is compact. Let $f: X \to Y$, $g: Y \to Z$ be continuous and injective, and $h = g \circ f$.
Prove that f is uniformly continuous if h is uniformly continuous. *Hint*: g^{-1} has compact domain $g(Y)$, and $f(x) = g^{-1}(h(x))$.
Prove also that f is continuous if h is continuous.

Exercise 17.12. Show that $f: [0, +\infty) \to [0, +\infty)$ defined by $f(x) = \sqrt{x}$ is uniformly continuous.

18 Differentiation

18.1 The Derivative of A Real Function

Definitions and Properties

The definition of the derivative of a real-valued function is no different from that which you have encountered in elementary calculus.

Definition 18.1 (Derivative). Suppose $f\colon [a,b] \to \mathbb{R}$. Let $x \in [a,b]$; if the limit

$$f'(x) := \lim_{t \to x} \frac{f(t)-f(x)}{t-x} \quad (a < t < b, t \neq x) \tag{18.1}$$

exists, we say f is ***differentiable*** at x; we call $f'(x)$ the ***derivative*** of f at x.
If f' is defined at every point of $E \subset [a,b]$, we say f is *differentiable on* E.
We also say f is *continuously differentiable* on E if f' exists at every point of E, and f' is continuous on E.

Equivalently, (18.1) can be written as

$$\lim_{h \to 0} \frac{f(x+h)-f(x)}{h} = f'(x),$$

or,

$$\frac{f(x+h)-f(x)}{h} = f'(x) + \varepsilon(h),$$

where $\varepsilon(h) \to 0$ as $h \to 0$. Rearranging gives

$$f(x+h) = f(x) + hf'(x) + h\varepsilon(h).$$

Using the small-o notation, we write $o(h)$ for a function that satisfies $o(h)/h \to 0$ as $h \to 0$. Hence we have

$$f(x+h) = f(x) + hf'(x) + o(h). \tag{18.2}$$

We can interpret (18.2) as an *approximation* of $f(x+h)$:

$$f(x+h) = \underbrace{f(x)+hf'(x)}_{\text{linear approximation}} + \underbrace{o(h)}_{\text{error term}}.$$

Lemma 18.2 (Differentiability implies continuity). *If $f\colon [a,b] \to \mathbb{R}$ is differentiable at $x \in [a,b]$, then f is continuous at x.*

Proof. Suppose $f\colon [a,b] \to \mathbb{R}$ is differentiable at $x \in [a,b]$. Then the limit $\lim\limits_{t\to x} \dfrac{f(t)-f(x)}{t-x}$ exists. Thus by the limits laws in 17.4,

$$\begin{aligned}\lim_{t\to x}[f(t)-f(x)] &= \lim_{t\to x}\left[\frac{f(t)-f(x)}{t-x}\cdot(t-x)\right] \\ &= \lim_{t\to x}\frac{f(t)-f(x)}{t-x}\cdot\lim_{t\to x}(t-x) \\ &= f'(x)\cdot 0 = 0.\end{aligned}$$

Since $\lim\limits_{t\to x} f(t) = f(x)$, by 17.10, f is continuous at x. □

Remark. The converse is not true; it is easy to construct continuous functions which fail to be differentiable at isolated points.

Example (Weierstrass function). Let $0 < a < 1$, let $b > 1$ be an odd integer, and $ab > 1 + \frac{3}{2}\pi$. Then the function

$$W(x) = \sum_{n=0}^{\infty} a^n \cos(b^n \pi x)$$

is continuous and nowhere differentiable on $\mathbb{R}$.

Example. One family of pathological examples in calculus is functions of the form

$$f(x) = x^p \sin\frac{1}{x}.$$

For $p = 1$, the function is continuous and differentiable everywhere other than $x = 0$; for $p = 2$, the function is differentiable everywhere, but the derivative is discontinuous.

Lemma 18.3 (Differentiation rules). *Suppose $f, g\colon [a,b] \to \mathbb{R}$ are differentiable at $x \in [a,b]$. Then*

(i) For a constant α, αf is differentiable at x, and *(scalar multiplication)*

$$(\alpha f)'(x) = \alpha f'(x).$$

(ii) $f+g$ is differentiable at x, and *(addition)*

$$(f+g)'(x) = f'(x) + g'(x).$$

(iii) fg is differentiable at x, and *(product rule)*

$$(fg)'(x) = f'(x)g(x) + f(x)g'(x).$$

(iv) f/g (when $g(x) \neq 0$) is differentiable at x, and *(quotient rule)*

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}.$$

Proof.

(i)

$$(\alpha f)'(x) = \lim_{t\to x} \frac{(\alpha f)(t) - (\alpha f)(x)}{t-x} = \alpha \lim_{t\to x} \frac{f(t)-f(x)}{t-x} = \alpha f'(x).$$

(ii)

$$\begin{aligned}
(f \pm g)'(x) &= \lim_{t\to x} \frac{(f+g)(t) - (f+g)(x)}{t-x} \\
&= \lim_{t\to x} \frac{f(t)+g(t)-f(x)-g(x)}{t-x} \\
&= \lim_{t\to x} \frac{f(t)-f(x)}{t-x} + \lim_{t\to x} \frac{g(t)-g(x)}{t-x} \\
&= f'(x) + g'(x)
\end{aligned}$$

(iii)

$$\begin{aligned}
(fg)'(x) &= \lim_{t\to x} \frac{(fg)(t) - (fg)(x)}{t-x} \\
&= \lim_{t\to x} \frac{f(t)g(t) - f(x)g(x)}{t-x} \\
&= \lim_{t\to x} \frac{[f(t)-f(x)]\,g(t) + f(x)\,[g(t)-g(x)]}{t-x} \\
&= \lim_{t\to x} \frac{f(t)-f(x)}{t-x} \cdot g(t) + \lim_{t\to x} f(x) \cdot \frac{g(t)-g(x)}{t-x} \\
&= f'(x)g(x) + f(x)g'(x)
\end{aligned}$$

(iv)

$$\begin{aligned}\left(\frac{f}{g}\right)'(x) &= \lim_{t\to x}\frac{\left(\frac{f}{g}\right)(t)-\left(\frac{f}{g}\right)(x)}{t-x}\\ &= \lim_{t\to x}\frac{\frac{f(t)}{g(t)}-\frac{f(x)}{g(x)}}{t-x}\\ &= \lim_{t\to x}\frac{1}{g(t)g(x)}\left[g(x)\cdot\frac{f(t)-f(x)}{t-x}-f(x)\cdot\frac{g(t)-g(x)}{t-x}\right]\\ &= \frac{f'(x)g(x)-f(x)g'(x)}{g(x)^2}\end{aligned}$$

□

By induction, we can obtain the following extensions of the differentiation rules.

Corollary. *Suppose $f_1, f_2, \ldots, f_n\colon [a,b]\to\mathbb{R}$ are differentiable at $x\in[a,b]$. Then*

(i) $f_1+f_2+\cdots+f_n$ is differentiable at x, and

$$(f_1+f_2+\cdots+f_n)'(x) = {f_1}'(x)+{f_2}'(x)+\cdots+{f_n}'(x).$$

(ii) $f_1f_2\cdots f_n$ is differentiable at x, and

$$(f_1f_2\cdots f_n)'(x) = {f_1}'(x)f_2(x)\cdots f_n(x)+f_1(x){f_2}'(x)\cdots f_n(x)+\cdots+f_1(x)f_2(x)\cdots {f_n}'(x).$$

The next result concerns the derivative of composition of functions.

Lemma 18.4 (Chain rule). *Suppose f is continuous on $[a,b]$, $f'(x)$ exists at $x\in[a,b]$, g is defined on I that contains $f([a,b])$, and g is differentiable at $f(x)$. Then $h=g\circ f$ is differentiable at x, and*

$$h'(x) = g'(f(x))\,f'(x). \tag{18.3}$$

Proof. By the definition of the derivative, we have

$$f(t)-f(x) = (t-x)[f'(x)+u(t)] \tag{1}$$

$$g(s)-g(f(x)) = (s-f(x))[g'(f(x))+v(s)] \tag{2}$$

where $t\in[a,b]$, $s\in I$, $\lim_{t\to x}u(t)=0$, $\lim_{s\to f(x)}v(s)=0$. ($u(t)$ and $v(s)$ can be viewed as some small

error terms which eventually go to 0.) Using first (2) and then (1), we obtain

$$\begin{aligned} h(t) - h(x) &= g(f(t)) - g(f(x)) \\ &= [f(t) - f(x)] \cdot [g'(f(x)) + v(s)] \\ &= (t-x)[f'(x) + u(t)][g'(f(x)) + v(s)], \end{aligned}$$

or, if $t \neq x$,

$$\frac{h(t) - h(x)}{t - x} = [g'(f(x)) + v(s)][f'(x) + u(t)].$$

Taking limits $t \to x$, we see that $u(t)$ and $v(s)$ eventually go to 0, so

$$h'(x) = \lim_{t \to x} \frac{h(t) - h(x)}{t - x} = g'(f(x))\, f'(x)$$

as desired. □

Later on when we talk about properties of differentiation such as the intermediate value theorems, we usually have the following requirement on the function:

f is continuous on $[a,b]$, differentiable on (a,b).

Derivatives of Higher Order

If f has a derivative f' on an interval, and if f' is itself differentiable, we denote the derivative of f' by f'', and call f'' the *second derivative* of f. Continuing in this manner, we obtain functions

$$f, f', f'', f^{(3)}, f^{(4)}, \ldots, f^{(n)},$$

each of which is the derivative of the preceding one. $f^{(n)}$ is called the n-th derivative (or the derivative or order n) of f.

Notation. $\mathscr{C}_1[a,b]$ denotes the set of differentiable functions over $[a,b]$ whose derivative is continuous. More generally, $\mathscr{C}_n[a,b]$ denotes the set of functions whose n-th derivative is continuous. In particular, $\mathscr{C}_0[a,b]$ is the set of continuous functions over $[a,b]$.

18.2 Mean Value Theorems

Let (X,d) be a metric space.

Definition 18.5. We say $f\colon X \to \mathbb{R}$ has

(i) a ***local maximum*** at $x \in X$ if there exists $\delta > 0$ such that $f(x) \geq f(t)$ for all $t \in B_\delta(x)$;

(ii) a ***local minimum*** at $x \in X$ if there exists $\delta > 0$ such that $f(x) \leq f(t)$ for all $t \in B_\delta(x)$.

Our next result is the basis of many applications of differentiation.

Lemma 18.6 (Fermat's theorem). *Suppose $f\colon [a,b] \to \mathbb{R}$. If f has a local maximum or minimum at $x \in (a,b)$, and if $f'(x)$ exists, then*

$$f'(x) = 0.$$

Proof. We prove the case for local maxima; the proof for the case for local minima is similar. Since x is a local maximum, choose $\delta > 0$ such that

$$a < x - \delta < x < x + \delta < b,$$

and $f(x) \geq f(t)$ for all $t \in (x-\delta, x+\delta)$. Then

$$\frac{f(t)-f(x)}{t-x} \begin{cases} \geq 0 & (x-\delta < t < x) \\ \leq 0 & (x < t < x+\delta) \end{cases}$$

Letting $t \to x$, we obtain

$$f'(x) \begin{cases} \geq 0 & (x-\delta < t < x) \\ \leq 0 & (x < t < x+\delta) \end{cases}$$

Hence we must have $f'(x) = 0$. □

Theorem 18.7 (Rolle's theorem). *Suppose f is continuous on $[a,b]$ and differentiable in (a,b). If $f(a) = f(b)$, then there exists $c \in (a,b)$ such that*

$$f'(c) = 0.$$

The idea is to show that f has a local maximum/minimum, then by Fermat's theorem this will then be the stationary point that we're trying to find.

Proof. Since f is continuous on $[a,b]$, by the extreme value theorem (17.19), f attains its maximum M and minimum m.

- If M and m both equal $f(a) = f(b)$, then f is simply a constant function; hence $f'(x) = 0$ for all $x \in [a,b]$.
- Otherwise, f has a maximum/minimum that does not equal $f(a) = f(b)$. Then there exists $c \in (a,b)$ such that $f(c)$ is a local maximum/minimum. Since f is differentiable on (a,b), $f'(c)$ exists, so by Fermat's theorem, $f'(c) = 0$.

□

Theorem 18.8 (Generalised mean value theorem)**.** *Suppose f and g are continuous on $[a,b]$ and differentiable in (a,b). Then there exists $c \in (a,b)$ such that*

$$[f(b) - f(a)]g'(c) = [g(b) - g(a)]f'(c). \tag{18.4}$$

Proof. For $t \in [a,b]$, define the *auxilliary function*

$$h(t) := [f(b) - f(a)]g(t) - [g(b) - g(a)]f(t).$$

Since f and g are continuous on $[a,b]$, this implies h is continuous on $[a,b]$; since f and g are differentiable on (a,b), this implies h is differentiable on (a,b). Moreover,

$$h(a) = f(b)g(a) - f(a)g(b) = h(b).$$

By Rolle's theorem, there exists $c \in (a,b)$ such that $h'(c) = 0$; that is,

$$[f(b) - f(a)]g'(c) = [g(b) - g(a)]f'(c)$$

as desired. □

Theorem 18.9 (Mean value theorem)**.** *Suppose f is continuous on $[a,b]$ and differentiable in (a,b). Then there exists $c \in (a,b)$ such that*

$$f(b) - f(a) = f'(c)(b-a). \tag{18.5}$$

Proof. Take $g(x) = x$ in 18.8. □

Lemma 18.10. *Suppose f is differentiable in (a,b).*

(i) If $f'(x) \geq 0$ for all $x \in (a,b)$, then f is monotonically increasing.

(ii) If $f'(x) = 0$ for all $x \in (a,b)$, then f is constant.

(iii) If $f'(x) \leq 0$ for all $x \in (a,b)$, then f is monotonically decreasing.

Proof. All conclusions can be read off from the equation

$$f'(x) = \frac{f(x_2) - f(x_1)}{x_2 - x_1},$$

which is valid, for each pair of numbers x_1, x_2 in (a,b), for some x between x_1 and x_2. □

18.3 A Restriction on Discontinuities of Derivatives

We have already seen that a function may have a derivative which exists at every point, but is discontinuous at some point. However, not every function is a derivative. In particular, derivatives which exist at every point of an interval have some sort of a "intermediate value" property – similar to continuous functions (see 17.24).

Theorem 18.11 (Darboux's theorem)**.** *Suppose f is differentiable on $[a,b]$, and suppose $f'(a) < c < f'(b)$. Then there exists $x \in (a,b)$ such that $f'(x) = c$.*

Proof. For $t \in (a,b)$, consider the auxilliary function

$$g(t) := f(t) - ct.$$

Then $g'(t) = f'(t) - c$. Thus

$$g'(a) = f'(a) - c < 0,$$

so there exists $t_1 \in (a,b)$ such that $g(t_1) < g(a)$. Similarly,

$$g'(b) = f'(b) - c > 0,$$

so there exists $t_2 \in (a,b)$ such that $g(t_2) < g(b)$.

By the extreme value theorem, g attains its minimum on $[a,b]$. From above, $g(a)$ and $g(b)$ cannot be minimums. Thus g must attain its minimum at some $x \in (a,b)$. By Fermat's theorem, $g'(x) = 0$. Hence $f'(x) = c$, as desired. □

Corollary 18.12. *If f is differentiable on $[a,b]$, then f' cannot have any simple discontinuities on $[a,b]$.*

Remark. But f' may very well have discontinuities of the second kind.

18.4 L'Hopital's Rule

An extremely useful result which arises from the generalised mean value theorem is known as *L'Hospital's rule*. It provides a simple procedure for the evaluation oflimiting values of functions which are expressible as quotients.

Lemma 18.13 (L'Hopital's rule). *Suppose f and g are differentiable over (a,b), with $g'(x) \neq 0$ for all $x \in (a,b)$, where $-\infty \leq a < b \leq +\infty$. If either*

(i) $\lim_{x\to a} f(x) = 0$ *and* $\lim_{x\to a} g(x) = 0$, *or* *(indeterminate $0/0$)*

(ii) $\lim_{x\to a} g(x) = +\infty$, *(indeterminate ∞/∞)*

and

$$\lim_{x\to a} \frac{f'(x)}{g'(x)} = A,$$

then

$$\lim_{x\to a} \frac{f(x)}{g(x)} = A.$$

The analogous statement is of course also true if $x > b$, or if $g(x) \to -\infty$ in (ii).

Proof. We first consider the case in which $-\infty \leq A < +\infty$. Choose $q \in \mathbb{R}$ such that $A < q$, and choose $r \in \mathbb{R}$ such that $A < r < q$. By (13) there exists $c \in (a,b)$ such that $a < x < c$ implies

$$\frac{f'(x)}{g'(x)} < r.$$

If $a < x < y < c$, then by the generalised mean value theorem (18.8), there exists $t \in (x,y)$ such that

$$\frac{f(x) - f(y)}{g(x) - g(y)} = \frac{f'(t)}{g'(t)} < r.$$

(i) Suppose $\lim_{x\to a} f(x) = 0$ and $\lim_{x\to a} g(x) = 0$. Let $x \to a$ in (18), we see that

$$\frac{f(y)}{g(y)} \leq r < q \quad (a < y < c).$$

(ii) Next, suppose $\lim_{x\to a} g(x) = +\infty$. Keeping y fixed in (18), we can choose a point $c_1 \in (a,y)$ such that $g(x) > g(y)$ and $g(x) > 0$ if $a < x < c_1$. Multiplying (18) by $[g(x) - g(y)]/g(x)$, we obtain

$$\frac{f(x)}{g(x)} < r - r\frac{g(y)}{g(x)} + \frac{f(y)}{g(x)} \quad (a < x < c_1).$$

If we let $x \to a$ in (20), (15) shows that there exists $c_2 \in (a, c_1)$ such that

$$\frac{f(x)}{g(x)} < q \quad (a < x < c_2).$$

Summing up, (19) and (21) show that for any q, subject only to the condition $A < q$, there is a point c_2 such that $f(x)/g(x) < q$ if $a < x < c_2$.

In the same manner, if $-\infty < A \leq +\infty$, and p is chosen so that $p < A$, we can find a point c_3 such that

$$p < \frac{f(x)}{g(x)} \quad (a < x < c_3),$$

and (16) follows from these two statements. □ to review proof

Example. To evaluate

$$\lim_{x\to 0} \frac{1-\cos x^2}{x^4}$$

we consider instead

$$\lim_{x\to 0} \frac{2x \sin x^2}{4x^3} = \lim_{x\to 0} \frac{\sin x^2}{2x^2} = \frac{1}{2}.$$

By the L'Hopital's rule, this is also the value of the original limit.

Example. Consider the limit

$$\lim_{x\to 0^+} x^x.$$

Since the exponential function is continuous, we conclude that

$$\lim_{x\to 0^+} x^x = \lim_{x\to 0^+} e^{x\log x} = e^{\lim\limits_{x\to 0^+} x\log x} = e^0 = 1$$

since

$$\lim_{x\to 0^+} x\log x = \lim_{x\to 0} \frac{\log x}{\frac{1}{x}} = \lim_{x\to 0^+} \frac{\frac{1}{x}}{-\frac{1}{x^2}} = \lim_{x\to 0^+} -x = 0.$$

18.5 Taylor's Theorem

Theorem 18.14 (Taylor's theorem). *Suppose $f\colon [a,b] \to \mathbb{R}$, $f^{(n-1)}$ is continuous on $[a,b]$, $f^{(n)}$ exists on (a,b). Assume that $c \in [a,b]$. Let the* Taylor polynomial *of degree $n-1$ of f at $x=c$ be*

$$\begin{aligned} P_{n-1}(x) &= \sum_{k=0}^{n-1} \frac{f^{(k)}(c)}{k!}(x-c)^k \\ &= f(c) + f'(c)(x-c) + \frac{f''(c)}{2!}(x-c)^2 + \cdots + \frac{f^{(n-1)}(c)}{(n-1)!}(x-c)^{n-1}. \end{aligned}$$

Then for every $x \in [a,b]$, $x \neq c$, there exists z_x between x and c such that

$$f(x) = P_{n-1}(x) + \frac{f^{(n)}(z_x)}{n!}(x-c)^n. \tag{18.6}$$

For $n=1$, this is just the mean value theorem. In general, the theorem shows that f can be approximated by a polynomial of degree $n-1$, and that (18.6) allows us to accurately estimate the error.

Proof. Let M be the number defined by

$$f(x) = P_{n-1}(x) + M(x-c)^n.$$

We claim that $n!M = f^{(n)}(z_x)$ for some z_x between x and c.

For all $x \in [a,b]$, let

$$g(x) = f(x) - P_{n-1}(x) - M(x-c)^n.$$

Then for all $x \in (a,b)$,

$$g^{(n)}(x) = f^{(n)}(x) - n!M.$$

Hence our proof will be complete if we can show that $g^{(n)}(z_x) = 0$ for some z_x between c and x. Since $P_{n-1}^{(k)}(c) = f^{(k)}(c)$ for $k = 0, \ldots, n-1$, we have

$$g(c) = g'(c) = \cdots = g^{(n-1)}(c) = 0.$$

By our choice of M, we have that $g(x) = 0$. By the mean value theorem, there exists x_1 between x and c such that $g'(x_1) = 0$. Since $g'(c) = 0$, we conclude similarly that $g''(x_2) = 0$ for some x_2 between x_1 and c. After n steps we arrive at the conclusion that $g^{(n)}(x_n) = 0$ for some x_n between x_{n-1} and c, that is, between x and c. □

18.6 Differentiation of Vector-valued Functions

Definition 5.1 applies without any change to complex functions f defined on $[a,b]$, and Theorems 5.2 and 5.3, as well as their proofs, remain valid. If f_1 and f_2 are the real and imaginary parts of f, that is, if

$$f(t) = f_1(t) + if_2(t)$$

for $a \leq t \leq b$, where $f_1(t)$ and $f_2(t)$ are real, then we clearly have

$$f'(x) = f_1'(x) + if_2'(x);$$

also, f is differentiable at x if and only if both f_1 and f_2 are differentiable at x.

Passing to vector-valued functions $\mathbf{f}$: $[a,b] \to \mathbb{R}^k$, we may still apply Definition 5.1 to define $\mathbf{f}'(x)$. The term $\phi(t)$ in (1) is now, for each t, a point in $\mathbb{R}^k$, and the limit in (2) is taken with respect to the norm of $\mathbb{R}^k$. In other words, $\mathbf{f}'(x)$ is that point of $\mathbb{R}^k$ (if there is one) for which

$$\lim_{t\to x}\left|\frac{\mathbf{f}(t)-\mathbf{f}(x)}{t-x} - \mathbf{f}'(x)\right| = 0,$$

and $\mathbf{f}'$ is again a function with values in $\mathbb{R}^k$.

If $f_1, \ldots, f_k$ are the components of $\mathbf{f}$, as defined in Theorem 4.10, then

$$\mathbf{f}' = (f_1, \ldots, f_k),$$

and $\mathbf{f}$ is differentiable at a point x if and only if each of the functions $f_1, \ldots, f_k$ is differentiable at x.

Theorem 5.2 is true in this context as well, and so is Theorem 5.3(a) and (b), if fg is replaced by the inner product $\mathbf{f} \cdot \mathbf{g}$ (see Definition 4.3).

When we turn to the mean value theorem, however, and to one of its consequences, namely, L'Hospital's rule, the situation changes. The next two examples will show that each of these results fails to be true for complex-valued functions.

Example. Define, for real x,

$$f(x) := e^{ix}.$$

Then $f(x) = \cos x + i \sin x$, so

$$f(2\pi) - f(0) = 1 - 1 = 0,$$

but $f'(x) = ie^{ix}$, so $|f'(x)| = 1$ for all real x.
Hence the mean value theorem fails to hold in this case.

Example. On $(0,1)$ define

$$f(x) := x, \quad g(x) := x + x^2 e^{i/x^2}.$$

Since $|e^{it} = 1|$, we see that

$$\lim_{x\to 0} \frac{f(x)}{g(x)} = 1.$$

Next,

$$g'(x) = 1 + \left(2x - \frac{2i}{x}\right) e^{i/x^2} \quad (0 < x < 1),$$

so that

$$|g'(x)| \geq \left|2x - \frac{2i}{x}\right| - 1 \geq \frac{2}{x} - 1.$$

Hence

$$\left|\frac{f'(x)}{g'(x)}\right| = \frac{1}{|g'(x)|} \leq \frac{x}{2-x}$$

and so

$$\lim_{x\to 0} \frac{f'(x)}{g'(x)} = 0.$$

By (36) and (40), L'Hospital's rule fails in this case. Note also that $g'(x) \neq 0$ on $(0,1)$, by (38).

However, there is a consequence of the mean value theorem which, for purposes of applications, is almost as useful as Theorem 5.10, and which remains true for vector-valued functions: From Theorem 5.10 it follows that

$$|f(b) - f(a)| \leq (b-a) \sup_{x\in[a,b]} |f'(x)|.$$

Theorem 18.15. *Suppose* $\mathbf{f}$: $[a,b] \to \mathbb{R}^k$ *is continuous on* $[a,b]$ *and differentiable in* (a,b). *Then there exists* $x \in (a,b)$ *such that*

$$\|\mathbf{f}(b) - \mathbf{f}(a)\| \leq (b-a)\|\mathbf{f}'(x)\|. \tag{18.7}$$

Proof. Put $\mathbf{z} = \mathbf{f}(b) - \mathbf{f}(a)$, and define

$$\phi(t) = \mathbf{z} \cdot \mathbf{f}(t) \quad (a \leq t \leq b).$$

Then ϕ is a real-valued continuous function on $[a,b]$ which is differentiable in (a,b). By the mean value theorem (18.9), there exists $x \in (a,b)$ such that

$$\phi(b) - \phi(a) = (b-a)\phi'(x) = (b-a)\mathbf{z} \cdot \mathbf{f}'(x).$$

On the other hand,

$$\begin{aligned}\phi(b)-\phi(a) &= \mathbf{z}\cdot\mathbf{f}(b)-\mathbf{z}\cdot\mathbf{f}(a)\\ &= \mathbf{z}\cdot(\mathbf{f}(b)-\mathbf{f}(a))\\ &= \mathbf{z}\cdot\mathbf{z} = \|\mathbf{z}\|^2.\end{aligned}$$

By the Cauchy–Schwarz inequality, we obtain

$$\|\mathbf{z}\|^2 = (b-a)\big\|\mathbf{z}\cdot\mathbf{f}'(x)\big\| \le (b-a)\|\mathbf{z}\|\,\big\|\mathbf{f}'(x)\big\|.$$

Hence $\|\mathbf{z}\| \le (b-a)\|\mathbf{f}'(x)\|$, which is the desired conclusion. □

Exercises

Exercise 18.1. Let f be monotonic on $[a,b]$, and differentiable with $f' \neq 0$ for $a < x < b$. Then f^{-1} is defined on an interval $[\alpha,\beta]$, and f^{-1} is differentiable in its interior, with

$$(f^{-1})'(\gamma) = \frac{1}{f'(f^{-1}(\gamma))}$$

for all $\gamma \in (\alpha,\beta)$.

Solution. We must show that

$$\lim_{x\to\gamma} \frac{f^{-1}(x) - f^{-1}(\gamma)}{x-\gamma}$$

exists. Set $f^{-1}(x) = y$ and $f^{-1}(\gamma) = c$. Then $x = f(y)$ and $\gamma = f(c)$, and since f and g are both continuous, the limit we must consider becomes

$$\lim_{y\to c} \frac{y-c}{f(y)-f(c)}$$

which we see at once to exist and be $\frac{1}{f'(c)} = \frac{1}{f'(f^{-1}(\gamma))}$. □

Exercise 18.2. Let f and g be continuous on $[a,b]$ and differentiable on (a,b). If $f'(x) = g'(x)$, then $f(x) = g(x) + C$.

Exercise 18.3. Given that $f(x) = x^\alpha$ where $0 < \alpha < 1$. Prove that f is uniformly continuous on $[0,+\infty)$.

Exercise 18.4. Let f be continuous on $[0,1]$ and differentiable on $(0,1)$ where $f(0) = f(1) = 0$. Prove that there exists $c \in (0,1)$ such that

$$f(x) + f'(x) = 0.$$

19 Riemann–Stieltjes Integral

The present chapter is based on a definition of the Riemann integral which depends very explicitly on the order structure of the real line. Accordingly, we begin by discussing integration of real-valued functions on intervals. Extensions to complex- and vector-valued functions on intervals follow in later sections.

19.1 Definition of Riemann–Stieltjes Integral

To approximate the area under the curve of a function, we partition the interval into finitely many sub-intervals, then multiply the width of each sub-interval by its height.

- For the height, we can choose to either use the supremum of the function over the interval or the infimum. Obviously, using the supremum will provide an upper bound, and using the infimum will provide a lower bound.
- For the width, we use the difference between the two endpoints in their output values when input into a monotonically increasing function α.

The upper Riemann integral is the infimum of upper bounds over all possible partitions. The lower Riemann integral is similarly defined. If they are equal, then the function is said to be Riemann–Stieltjes integrable.

Notation and Preliminaries

A ***partition*** P of a closed interval $[a,b]$ is a finite set of points $\{x_0, x_1, \ldots, x_n\}$, where

$$a = x_0 \leq x_1 \leq \cdots \leq x_{n-1} \leq x_n = b.$$

Notation. Denote the set of all partitions of $[a,b]$ by $\mathscr{P}[a,b]$.

Let $f\colon [a,b] \to \mathbb{R}$ be bounded. Denote

$$M_i = \sup_{x \in [x_{i-1}, x_i]} f(x), \quad m_i = \inf_{x \in [x_{i-1}, x_i]} f(x) \quad (i = 1, \ldots, n).$$

Let α be a monotonically increasing function on $[a,b]$. Denote

$$\Delta\alpha_i = \alpha(x_i) - \alpha(x_{i-1}) \quad (i = 1,\dots,n).$$

(These suprema and infima are well-defined, finite real numbers due to the boundedness of f.)

The *upper sum* and *lower sum* of f with respect to the partition P and α are respectively

$$U(f,\alpha;P) = \sum_{i=1}^{n} M_i \Delta\alpha_i,$$
$$L(f,\alpha;P) = \sum_{i=1}^{n} m_i \Delta\alpha_i.$$

insert diagram

The partition P' is a ***refinement*** of P if $P' \supset P$. Given two partitions P_1 and P_2, we say that P' is their *common refinement* if $P' = P_1 \cup P_2$.

Intuitively, a refinement will give a better estimation than the original partition, so the upper and lower sums of a refinement should be more restrictive.

Lemma 19.1. *If P' is a refinement of P, then*

(i) $L(f,\alpha;P) \le L(f,\alpha;P')$

(ii) $U(f,\alpha;P') \le U(f,\alpha;P)$

Proof.

(i) Suppose first that P' contains just one point more than P. Let this extra point be x', and suppose $x_{i-1} < x' < x_i$ for some i $(1 \le i \le n)$, where $x_{i-1}, x_i \in P$. Let

$$w_1 = \inf_{x\in[x_{i-1},x']} f(x), \quad w_2 = \inf_{x\in[x',x_i]} f(x).$$

Let, as before,

$$m_i = \inf_{x\in[x_{i-1},x_i]} f(x).$$

Clearly $w_1 \ge m_i$ and $w_2 \ge m_i$. Then

$$\begin{aligned}
&L(f,\alpha;P') - L(f,\alpha;P) \\
&= w_1\big(\alpha(x') - \alpha(x_{i-1})\big) + w_2\big(\alpha(x_i) - \alpha(x')\big) - m_i\big(\alpha(x_i) - \alpha(x_{i-1})\big) \\
&= \underbrace{(w_1 - m_i)}_{\ge 0}\underbrace{\big(\alpha(x') - \alpha(x_{i-1})\big)}_{>0} + \underbrace{(w_2 - m_i)}_{\ge 0}\underbrace{\big(\alpha(x_i) - \alpha(x')\big)}_{>0} \\
&\ge 0
\end{aligned}$$

and hence $L(f,\alpha;P) \le L(f,\alpha;P')$.

If P' contains k more points than P, we repeat this reasoning k times.

(ii) Analogous to the proof of (i).

□

Since f is bounded, there exist m and M such that $m \leq f(x) \leq M$ for all $x \in [a,b]$. Hence for every partition P,

$$m(\alpha(b)-\alpha(a)) \leq L(f,\alpha;P) \leq U(f,\alpha;P) \leq M(\alpha(b)-\alpha(a))$$

so that the numbers $L(f,\alpha;P)$ and $U(f,\alpha;P)$ form a bounded set. This shows that the upper and lower integrals are defined for every bounded function f. We now define the *upper and lower Riemann–Stieltjes integrals* respectively as

$$\overline{\int_a^b} f\,\mathrm{d}\alpha := \inf_{P\in\mathscr{P}[a,b]} U(f,\alpha;P)$$
$$\underline{\int_a^b} f\,\mathrm{d}\alpha := \sup_{P\in\mathscr{P}[a,b]} L(f,\alpha;P)$$

where we take inf and sup over all partitions.

One would expect the lower RS integral to be less than or equal to the upper RS integral. We now show this.

Lemma 19.2.

$$\underline{\int_a^b} f\,\mathrm{d}\alpha \leq \overline{\int_a^b} f\,\mathrm{d}\alpha\,.$$

Proof. Let P' be the common refinement of partitions P_1 and P_2; that is, $P' = P_1 \cup P_2$. Clearly $P' \supset P_1$; by 19.1,

$$L(f,\alpha;P_1) \leq L(f,\alpha;P').$$

Similarly, $P' \supset P_2$, so

$$U(f,\alpha;P') \leq U(f,\alpha;P_2).$$

Clearly $L(f,\alpha;P') \leq U(f,\alpha;P')$. Thus combining the above two equations gives

$$L(f,\alpha;P_1) \leq U(f,\alpha;P_2).$$

Fix P_2 and take sup over all P_1 gives

$$\underline{\int_a^b} f\,\mathrm{d}\alpha \leq U(f,\alpha;P_2).$$

Then taking inf over all P_2 gives

$$\underline{\int_a^b} f \, \mathrm{d}\alpha \leq \overline{\int_a^b} f \, \mathrm{d}\alpha \, .$$

□

Defining the Integral

Definition 19.3 (Riemann–Stieltjes integral). We say $f : [a,b] \to \mathbb{R}$ is ***Riemann–Stieltjes integrable*** with respect to α over $[a,b]$, if

$$\underline{\int_a^b} f \,\mathrm{d}\alpha = \overline{\int_a^b} f \,\mathrm{d}\alpha .$$

We call the common value the ***Riemann–Stieltjes integral*** of f with respect to α over $[a,b]$, and denote it as

$$\int_a^b f \,\mathrm{d}\alpha .$$

The functions f and α are referred to as the *integrand* and the *integrator*, respectively.

Notation. $\mathscr{R}(\alpha)$ denotes the set of Riemann–Stieltjes integrable functions with respect to α.

In particular, when $\alpha(x) = x$, we call the corresponding Riemann–Stieltjes integral the *Riemann integral*, and use $\mathscr{R}$ to denote the set of Riemann integrable functions.

Notation. Since x is a "dummy variable" and may be replaced by any other variable, we shall omit it.

Example (Dirichlet function). The *Dirichlet function* is defined over $[0,1]$ by

$$f(x) = \begin{cases} 1 & (x \in \mathbb{Q}) \\ 0 & (x \in \mathbb{R} \setminus \mathbb{Q}) \end{cases}$$

For each subinterval $[x_{i-1}, x_i]$, due to the density of rationals and irrationals, $[x_{i-1}, x_i]$ contains both rationals and irrationals, so $M_i = 1$ and $m_i = 0$. Thus for any partition P,

$$U(f;P) = 1, \quad L(f;P) = 0.$$

Therefore,

$$1 = \overline{\int} f \,\mathrm{d}\alpha \neq \underline{\int} f \,\mathrm{d}\alpha = 0$$

so the Dirichlet function is not Riemann–Stieltjes integrable.

The next result is particularly useful in determining the Riemann–Stieltjes integrability of a function. We will use it many times later.

Lemma 19.4 (Integrability criterion). *$f \in \mathscr{R}(\alpha)$ if and only if*

$$\forall \varepsilon > 0, \quad \exists P, \quad U(f,\alpha;P) - L(f,\alpha;P) < \varepsilon.$$

Proof.

$\boxed{\Longrightarrow}$ Suppose $f \in \mathscr{R}(\alpha)$. Let $\varepsilon > 0$ be given. Then there exists partitions P_1 and P_2 such that

$$U(f,\alpha;P_2) - \int_a^b f\,\mathrm{d}\alpha < \frac{\varepsilon}{2}$$

and

$$\int_a^b f\,\mathrm{d}\alpha - L(f,\alpha;P_1) < \frac{\varepsilon}{2}.$$

Choose P to be the common refinement of P_1 and P_2. Then

$$\begin{aligned} U(f,\alpha;P) &\le U(f,\alpha;P_2) \\ &< \int_a^b f\,\mathrm{d}\alpha + \frac{\varepsilon}{2} \\ &< L(f,\alpha;P_1) + \varepsilon \\ &\le L(f,\alpha;P) + \varepsilon. \end{aligned}$$

Hence for this partition P, we have

$$U(f,\alpha;P) - L(f,\alpha;P) < \varepsilon.$$

$\boxed{\Longleftarrow}$ By 19.2, for every partition P,

$$L(f,\alpha;P) \le \underline{\int_a^b} f\,\mathrm{d}\alpha \le \overline{\int_a^b} f\,\mathrm{d}\alpha \le U(f,\alpha;P).$$

Since $U(f,\alpha;P) - L(f,\alpha;P) < \varepsilon$, we have

$$0 \le \overline{\int_a^b}\,\mathrm{d}\alpha - \underline{\int_a^b} f\,\mathrm{d}\alpha < \varepsilon.$$

Since this holds for all $\varepsilon > 0$, we have

$$\overline{\int_a^b} f\,\mathrm{d}\alpha = \underline{\int_a^b} f\,\mathrm{d}\alpha\,.$$

Hence $f \in \mathscr{R}(\alpha)$. □

Useful Identities

Proposition 19.5 (Cauchy criterion).

(i) If $U(f,\alpha;P)-L(f,\alpha;P)<\varepsilon$ holds for some P and some $\varepsilon>0$, then $U(f,\alpha;P')-L(f,\alpha;P')<\varepsilon$ holds (with the same ε) for every refinement of P, P'.

(ii) If $U(f,\alpha;P)-L(f,\alpha;P)<\varepsilon$ holds for $P=\{x_0,\dots,x_n\}$, and

$$s_i,t_i\in[x_{i-1},x_i]\quad(i=1,\dots,n)$$

then

$$\sum_{i=1}^{n}|f(s_i)-f(t_i)|\,\Delta\alpha_i<\varepsilon.$$

(iii) If $f\in\mathscr{R}(\alpha)$ and the hypotheses of (ii) hold, then

$$\left|\sum_{i=1}^{n}f(t_i)\Delta\alpha_i-\int_a^b f\,d\alpha\right|<\varepsilon.$$

Proof.

(i) Suppose $U(f,\alpha;P)-L(f,\alpha;P)<\varepsilon$ holds for some partition P and some $\varepsilon>0$. By 19.1, for any refinement P',

$$U(f,\alpha;P')\le U(f,\alpha;P),\quad L(f,\alpha;P)\le L(f,\alpha;P').$$

Hence

$$U(f,\alpha;P')-L(f,\alpha;P')\le U(f,\alpha;P)-L(f,\alpha;P)<\varepsilon.$$

(ii) Since

$$f(s_i),f(t_i)\in[m_i,M_i]\quad(i=1,\dots,n)$$

it follows that

$$|f(s_i)-f(t_i)|\le M_i-m_i.$$

Hence

$$\sum_{i=1}^{n}|f(s_i)-f(t_i)|\,\Delta\alpha_i\le U(f,\alpha;P)-L(f,\alpha;P)<\varepsilon.$$

(iii) The desired result follows from the two inequalities

$$L(f,\alpha;P)\le\sum_{i=1}^{n}f(t_i)\Delta\alpha_i\le U(f,\alpha;P)$$

$$L(f,\alpha;P) \le \int_a^b f\,\mathrm{d}\alpha \le U(f,\alpha;P)$$

□

The next result states that all continuous functions are integrable.

Proposition 19.6 (Continuity implies integrability). *If f is continuous on $[a,b]$, then $f \in \mathscr{R}(\alpha)$.*

Proof. Let $\varepsilon > 0$ be given. Choose $\eta > 0$ such that

$$(\alpha(b) - \alpha(a))\,\eta < \varepsilon.$$

Since f is continuous on $[a,b]$ which is compact, by 17.28, f is uniformly continuous on $[a,b]$. Thus there exists $\delta > 0$ such that for all $x, y \in [a,b]$,

$$|x-y| < \delta \implies |f(x) - f(y)| < \eta.$$

If P is any partition of $[a,b]$ such that $\Delta x_i < \delta$ for $i = 1,\dots,n$, then

$$M_i - m_i \le \eta \quad (i = 1,\dots,n).$$

Hence

$$\begin{aligned} U(f,\alpha;P) - L(f,\alpha;P) &= \sum_{i-1}^{n} (M_i - m_i)\Delta\alpha_i \\ &\le \eta \sum_{i=1}^{n} \Delta\alpha_i = \eta\,(\alpha(b) - \alpha(a)) < \varepsilon. \end{aligned}$$

Therefore $f \in \mathscr{R}(\alpha)$, by the integrability criterion (19.4). □

Proposition 19.7. *If f is monotonic on $[a,b]$, and if α is continuous on $[a,b]$, then $f \in \mathscr{R}(\alpha)$.*

Proof. Let $\varepsilon > 0$ be given. For any positive integer n, choose a partition P such that

$$\Delta\alpha_i = \frac{\alpha(b) - \alpha(a)}{n} \quad (i = 1,\dots,n).$$

This is possible by the intermediate value theorem, due to the continuity of α.

Suppose that f is monotonically increasing (the proof is analogous in the other case). Then

$$M_i = f(x_i), \quad m_i = f(x_{i-1}) \quad (i = 1,\dots,n).$$

Hence

$$\begin{aligned} U(f,\alpha;P) - L(f,\alpha;P) &= \sum_{i=1}^{n} (M_i - m_i)\Delta\alpha_i \\ &= \frac{\alpha(b)-\alpha(a)}{n} \sum_{i=1}^{n} (f(x_i) - f(x_{i-1})) \\ &= \frac{\alpha(b)-\alpha(a)}{n} (f(b) - f(a)) < \varepsilon \end{aligned}$$

if n is taken large enough. Hence $f \in \mathscr{R}(\alpha)$, by the integrability criterion. □

Proposition 19.8. *Suppose f is bounded on $[a,b]$, f has only finitely many points of discontinuity on $[a,b]$, and α is continuous at every point at which f is discontinuous. Then $f \in \mathscr{R}(\alpha)$.*

Proof. Let $\varepsilon > 0$ be given. Since f is bounded, let $M = \sup |f(x)|$. Let E be the set of points at which f is discontinuous.

Since E is finite, and α is continuous at every point of E, we can cover E by finitely many disjoint intervals $[u_j, v_j] \subset [a,b]$ such that the sum of the corresponding differences $\sum_j \big(\alpha(v_j) - \alpha(u_j)\big) < \varepsilon$. Furthermore, we can place these intervals in such a way that every point of $E \cap (a,b)$ lies in the interior of some $[u_j, v_j]$.

Remove the segments (u_j, v_j) from $[a,b]$. The remaining set K is compact. Hence f is uniformly continuous on K, so there exists $\delta > 0$ such that for all $s,t \in K$,

$$|s-t| < \delta \implies |f(x) - f(t)| < \varepsilon.$$

Now form a partition $P = \{x_0, x_1, \ldots, x_n\}$ of $[a,b]$ as follows: Each u_j occurs in P. Each v_j occurs in P. No point of any segment (u_j, v_j) occurs in P. If x_{i-i} is not one of the u_j, then $\Delta x_i < \delta$.

Note that $M_i - m_i \leq 2M$ for every i, and that $M_i - m_i < \varepsilon$ unless x_{i-i} is one of the u_j. Hence

$$\begin{aligned} U(f,\alpha;P) - L(f,\alpha;P) &= \sum_{i=1}^{n} (M_i - m_i)\Delta\alpha_i \\ &\leq (\alpha(b) - \alpha(a))\,\varepsilon + 2M\varepsilon. \end{aligned}$$

Since ε is arbitrary, we have $f \in \mathscr{R}(\alpha)$, by the integrability criterion. □

The next result states that a uniformly continuous function of an integrable function is also integrable.

Proposition 19.9. *Suppose $f \in \mathscr{R}(\alpha)$, $m \le f \le M$, and ϕ is continuous on $[m,M]$. Then $\phi \circ f \in \mathscr{R}(\alpha)$.*

Proof. Let $h = \phi \circ f$. Let $\varepsilon > 0$ be given. Since ϕ is uniformly continuous on $[m,M]$, there exists $\delta > 0$ such that $\delta < \varepsilon$, and for all $s,t \in [m,M]$,

$$|s-t| \le \delta \implies |\phi(s) - \phi(t)| < \varepsilon.$$

Since $f \in \mathscr{R}(\alpha)$, by 19.4, there exists a partition $P = \{x_0, \ldots, x_n\}$ of $[a,b]$ such that

$$U(f,\alpha;P) - L(f,\alpha;P) < \delta^2. \tag{1}$$

Let

$$M_i = \sup_{x \in [x_{i-1},x_i]} f(x), \quad M_i^* = \sup_{x \in [x_{i-1},x_i]} h(x),$$
$$m_i = \inf_{x \in [x_{i-1},x_i]} f(x), \quad m_i^* = \inf_{x \in [x_{i-1},x_i]} h(x).$$

Divide the numbers $1, \ldots, n$ into two classes:

$$A = \{i \mid M_i - m_i < \delta\},$$
$$B = \{i \mid M_i - m_i \ge \delta\}.$$

- For $i \in A$, our choice of δ shows that $M_i^* - m_i^* \le \varepsilon$.
- For $i \in B$, $M_i^* - m_i^* \le 2K$, where $K = \sup_{m \le t \le M} |\phi(t)|$.

By (1), we have

$$\delta \sum_{i \in B} \Delta\alpha_i \le \sum_{i \in B} (M_i - m_i)\Delta\alpha_i < \delta^2$$

so that $\sum_{i \in B} \Delta\alpha_i < \delta$. It follows that

$$\begin{aligned} U(h,\alpha;P) - L(h,\alpha;P) &= \sum_{i \in A} (M_i^* - m_i^*)\Delta\alpha_i + \sum_{i \in B} (M_i^* - m_i^*)\Delta\alpha_i \\ &\le \varepsilon\,(\alpha(b) - \alpha(a)) + 2K\delta \\ &< \varepsilon\,(\alpha(b) - \alpha(a) + 2K)\,. \end{aligned}$$

Since ε was arbitrary, by the integrability criterion, $h \in \mathscr{R}(\alpha)$. □

19.2 Properties of the Integral

Lemma 19.10.

(i) If $f_1, f_2 \in \mathscr{R}(\alpha)$, then $f_1 + f_2 \in \mathscr{R}(\alpha)$, and

$$\int_a^b (f_1 + f_2)\,\mathrm{d}\alpha = \int_a^b f_1\,\mathrm{d}\alpha + \int_a^b f_2\,\mathrm{d}\alpha\,.$$

(ii) If $f \in \mathscr{R}(\alpha)$, then $cf \in \mathscr{R}(\alpha)$ for every $c \in \mathbb{R}$, and

$$\int_a^b (cf)\,\mathrm{d}\alpha = c\int_a^b f\,\mathrm{d}\alpha\,.$$

(iii) If $f_1, f_2 \in \mathscr{R}(\alpha)$ and $f_1 \le f_2$, then

$$\int_a^b f_1\,\mathrm{d}\alpha \le \int_a^b f_2\,\mathrm{d}\alpha\,.$$

(iv) If $f \in \mathscr{R}(\alpha)$ and $c \in [a,b]$, then $f \in \mathscr{R}_\alpha[a,c]$ and $f \in \mathscr{R}_\alpha[c,b]$, and

$$\int_a^b f\,\mathrm{d}\alpha = \int_a^c f\,\mathrm{d}\alpha + \int_c^b f\,\mathrm{d}\alpha\,.$$

(v) If $f \in \mathscr{R}(\alpha)$ and $|f| \le M$, then

$$\left|\int_a^b f\,\mathrm{d}\alpha\right| \le M\,(\alpha(b) - \alpha(a))\,.$$

(vi) If $f \in R_{\alpha_1}[a,b]$ and $f \in R_{\alpha_2}[a,b]$, then $f \in R_{\alpha_1+\alpha_2}[a,b]$, and

$$\int_a^b f\,\mathrm{d}(\alpha_1 + \alpha_2) = \int_a^b f\,\mathrm{d}\alpha_1 + \int_a^b f\,\mathrm{d}\alpha_2\,;$$

if $f \in \mathscr{R}(\alpha)$ and c is a positive constant, then $f \in R_{c\alpha}[a,b]$, and

$$\int_a^b f\,\mathrm{d}(c\alpha) = c\int_a^b f\,\mathrm{d}\alpha\,.$$

(vii) If $f \in \mathscr{R}(\alpha)$ and $g \in \mathscr{R}(\alpha)$, then $fg \in \mathscr{R}(\alpha)$.

Proof.

(i) If $f = f_1 + f_2$ and P is any partition of $[a,b]$, we have

$$L(f_1,\alpha;P)+L(f_2,\alpha;P) \le L(f,\alpha;P) \le U(f,\alpha;P) \le U(f_1,\alpha;P)+U(f_2,\alpha;P). \quad (1)$$

If $f_1 \in \mathscr{R}(\alpha)$ and $f_2 \in \mathscr{R}(\alpha)$, let $\varepsilon > 0$ be given. There are partitions P_1 and P_2 such that

$$U(f_1,\alpha;P_1) - L(f_1,\alpha;P_1) < \frac{\varepsilon}{2}$$
$$U(f_2,\alpha;P_2) - L(f_2,\alpha;P_2) < \frac{\varepsilon}{2}$$

Let P be the common refinement of P_1 and P_2. Then (1) implies

$$U(f,\alpha;P) - L(f,\alpha;P) < \varepsilon$$

which proves that $f \in \mathscr{R}(\alpha)$.

With this same P we have

$$U(f_1,\alpha;P) < \int_a^b f_1 \,\mathrm{d}\alpha + \frac{\varepsilon}{2}$$
$$U(f_2,\alpha;P) < \int_a^b f_2 \,\mathrm{d}\alpha + \frac{\varepsilon}{2}$$

Hence (1) implies

$$\int_a^b f \,\mathrm{d}\alpha \le U(f,\alpha;P) < \int_a^b f_1 \,\mathrm{d}\alpha + \int_a^b f_2 \,\mathrm{d}\alpha + \varepsilon.$$

Since ε was arbitrary, we conclude that

$$\int_a^b f \,\mathrm{d}\alpha \le \int_a^b f_1 \,\mathrm{d}\alpha + \int_a^b f_2 \,\mathrm{d}\alpha\,.$$

If we replace f_1 and f_2 in the above equation by $-f_1$ and $-f_2$, the inequality is reversed, and the equality is proved.

(ii) The case where $c = 0$ is trivial. Given $\varepsilon > 0$, there exists P such that $U(f,\alpha;P) - L(f,\alpha;P) < \varepsilon$. If $c > 0$ write

$$U(cf,\alpha;P) = \sum_{i=1}^n cM_i\alpha_i = c\sum_{i=1}^n M_i\alpha_i = cU(f,\alpha;P).$$

Similarly,

$$L(cf,\alpha;P) = cL(f,\alpha;P).$$

Then

$$U(cf,\alpha;P) - L(cf,\alpha;P) = c\,(U(f,\alpha;P) - L(f,\alpha;P)) < c\varepsilon$$

and since ε is arbitrary, we are done. The case where $c < 0$ is similar. Therefore $cf \in \mathscr{R}(\alpha)$.

With this same P we have

$$U(f,\alpha;P) - \int_a^b f\,\mathrm{d}\alpha < \varepsilon.$$

Then if $c > 0$,

$$\int_a^b cf\,\mathrm{d}\alpha \le U(cf,\alpha;P) = cU(f,\alpha;P) < c\int_a^b f\,\mathrm{d}\alpha + c\varepsilon$$

so

$$\int_a^b cf\,\mathrm{d}\alpha \le c\int_a^b f\,\mathrm{d}\alpha\,.$$

If we replace f in the above equation by $-f$, the inequality is reversed, and the equality is proved.

(iii) For every partition P, we have

$$U(f_1,\alpha;P) = \sum_{i=1}^n M_i(f_1)\Delta\alpha_i \le \sum_{i=1}^n M_i(f_2)\Delta\alpha_i = U(f_2,\alpha;P)$$

since α is monotonically increasing on $[a,b]$.

(iv)

(v)

(vi)

(vii) Take $\phi(t) = t^2$. By 19.9, $f^2 \in R_\alpha[a,b]$ if $f \in R_\alpha[a,b]$. Write

$$fg = \frac{1}{4}\left((f+g)^2 - (f-g)^2\right).$$

Then the desired result follows.

□

Lemma 19.11 (Triangle inequality). *Suppose $f \in \mathscr{R}(\alpha)$. Then $|f| \in \mathscr{R}(\alpha)$, and*

$$\left|\int_a^b f\,\mathrm{d}\alpha\right| \le \int_a^b |f|\,\mathrm{d}\alpha\,.$$

Proof. Take $\phi(t) = |t|$, which is a continuous function. By 19.9, we have that $|f| = \phi \circ f \in \mathscr{R}(\alpha)$. Choose $c = \pm 1$, so that

$$c\int_a^b f\,\mathrm{d}\alpha \ge 0.$$

Then

$$\left|\int_a^b f\,\mathrm{d}\alpha\right| = c\int_a^b f\,\mathrm{d}\alpha = \int_a^b cf\,\mathrm{d}\alpha \le \int_a^b |f|\,\mathrm{d}\alpha,$$

since $cf \le |f|$. □

Example (Heaviside step function)**.** The *Heaviside step function* is defined by

$$H(x) = \begin{cases} 0 & (x \le 0) \\ 1 & (x > 0) \end{cases}$$

Proposition. *Suppose f is bounded on $[a,b]$, continuous at $s \in (a,b)$. Let $\alpha(x) = H(x-s)$, then*

$$\int_a^b f\,\mathrm{d}\alpha = f(s).$$

Proof. Consider partitions $P = \{x_0, x_1, x_2, x_3\}$, where $x_0 = a$, and $x_1 = s < x_2 < x_3 = b$. Then

$$U(f,\alpha;P) = M_2, \quad L(f,\alpha;P) = m_2.$$

Since f is continuous at s, we see that M_2 and m_2 converge to $f(s)$ as $x_2 \to s$. □

Proposition. *Suppose $c_n \ge 0$ for $n = 1,2,\ldots$, $\sum c_n$ converges, (s_n) is a sequence of distinct points in (a,b), and*

$$\alpha(x) = \sum_{n=1}^{\infty} c_n H(x-s_n).$$

Let f be continuous on $[a,b]$. Then

$$\int_a^b f\,\mathrm{d}\alpha = \sum_{n=1}^{\infty} c_n f(s_n).$$

Proof. Since $0 \le c_n H(x-s_n) \le c_n$ for $n = 1,2,\ldots$ and $\sum c_n$ converges, by the comparison test, $\alpha(x) = \sum c_n H(x-s_n)$ converges for every x. Its sum $\alpha(x)$ is evidently monotonic (since each term in the sum is non-negative), and $\alpha(a) = 0$, $\alpha(b) = \sum c_n$. Let $\varepsilon > 0$ be given. Since $\sum c_n$ converges, choose $N \in \mathbb{N}$ so that

$$\sum_{n=N+1}^{\infty} c_n < \varepsilon.$$

Let

$$\alpha_1(x) = \sum_{n=1}^{N} c_n H(x-s_n), \quad \alpha_2(x) = \sum_{n=N+1}^{\infty} c_n H(x-s_n).$$

By the previous result,

$$\int_a^b f\,\mathrm{d}\alpha_1 = \sum_{n=1}^{N} c_n f(s_n).$$

Since $\alpha_2(b) - \alpha_2(a) < \varepsilon$,

$$\left| \int_a^b f\,\mathrm{d}\alpha_2 \right| \leq M\varepsilon,$$

where $M = \sup |f(x)|$. Since $\alpha = \alpha_1 + \alpha_2$,

$$\int_a^b f\,\mathrm{d}\alpha = \int_a^b f\,\mathrm{d}\alpha_1 + \int_a^b f\,\mathrm{d}\alpha_2$$

so it follows that

$$\left| \int_a^b f\,\mathrm{d}\alpha - \sum_{n=1}^{N} c_n f(s_n) \right| \leq M\varepsilon.$$

Since ε was arbitrary, and taking $N \to \infty$, we obtain

$$\int_a^b f\,\mathrm{d}\alpha = \sum_{n=1}^{\infty} c_n f(s_n).$$

□

In this case, we call $\alpha(x)$ a *step function*; then the integral reduces to a finite or infinite series.

The next result states that if α has an integrable derivative, then the integral reduces to an ordinary Riemann integral.

Proposition 19.12. *Assume α increases monotonically, $\alpha' \in \mathscr{R}$. Let $f\colon [a,b] \to \mathbb{R}$ be bounded, then $f \in \mathscr{R}(\alpha)$ if and only if $f\alpha' \in \mathscr{R}$. In that case*

$$\int_a^b f\,\mathrm{d}\alpha = \int_a^b f(x)\alpha'(x)\,\mathrm{d}x. \tag{19.1}$$

Proof. Let $\varepsilon > 0$ be given and apply 19.4 to α': There exists a partition $P = \{x_0, \ldots, x_n\}$ of $[a,b]$ such that

$$U(\alpha'; P) - L(\alpha'; P) < \varepsilon. \tag{1}$$

By the mean value theorem, there exist points $t_i \in [x_{i-1}, x_i]$ such that

$$\Delta\alpha_i = \alpha'(t_i)\Delta x_i \quad (i = 1, \ldots, n).$$

If $s_i \in [x_{i-1}, x_i]$, then by 19.5,

$$\sum_{i=1}^{n} \left|\alpha'(s_i) - \alpha'(t_i)\right| \Delta x_i < \varepsilon. \tag{2}$$

Let $M = \sup|f(x)|$. Since

$$\sum_{i=1}^{n} f(s_i)\Delta\alpha_i = \sum_{i=1}^{n} f(s_i)\alpha'(t_i)\Delta x_i$$

it follows from (2) that

$$\begin{aligned}
\left|\sum_{i=1}^{n} f(s_i)\Delta\alpha_i - \sum_{i=1}^{n} f(s_i)\alpha'(s_i)\Delta x_i\right| &= \left|\sum_{i=1}^{n} f(s_i)\left(\alpha'(t_i) - \alpha'(s_i)\right)\Delta x_i\right| \\
&\le \sum_{i=1}^{n} \left|f(s_i)\left(\alpha'(t_i) - \alpha'(s_i)\right)\Delta x_i\right| \\
&= \sum_{i=1}^{n} |f(s_i)|\left|\alpha'(t_i) - \alpha'(s_i)\right|\Delta x_i \\
&\le M\sum_{i=1}^{n} \left|\alpha'(t_i) - \alpha'(s_i)\right|\Delta x_i \\
&\le M\varepsilon. \qquad (3)
\end{aligned}$$

In particular, for all choices of $s_i \in [x_{i-1}, x_i]$,

$$\sum_{i=1}^{n} f(s_i)\Delta\alpha_i \le U(f\alpha'; P) + M\varepsilon$$

so taking sup for $f(s_i)$ gives

$$U(f, \alpha; P) \le U(f\alpha'; P) + M\varepsilon.$$

The same argument leads from (3) to

$$U(f\alpha'; P) \le U(f, \alpha; P) + M\varepsilon.$$

Hence

$$\left|U(f, \alpha; P) - U(f\alpha'; P)\right| \le M\varepsilon. \tag{4}$$

Since (1) holds true for any refinement of P, hence (4) also remains true. We conclude that

$$\left|\overline{\int_a^b} f\,\mathrm{d}\alpha - \overline{\int_a^b} f(x)\alpha'(x)\,\mathrm{d}x\right| \le M\varepsilon.$$

But ε is arbitrary. Hence

$$\overline{\int_a^b} f\,\mathrm{d}\alpha = \overline{\int_a^b} f(x)\alpha'(x)\,\mathrm{d}x$$

for any bounded f. The equality of the lower integrals follows from

$$\begin{aligned}\overline{\int_a^b} -f\,\mathrm{d}\alpha &= \overline{\int_a^b} -f\alpha'\,\mathrm{d}x\\ -\underline{\int_a^b} f\,\mathrm{d}\alpha &= -\underline{\int_a^b} f\alpha'\,\mathrm{d}x\\ \underline{\int_a^b} f\,\mathrm{d}\alpha &= \underline{\int_a^b} f(x)\alpha'(x)\,\mathrm{d}x\end{aligned}$$

Therefore the theorem follows. □

Proposition 19.13 (Change of variables). *Suppose $\phi\colon [A,B]\to[a,b]$ is strictly increasing and continuous. Suppose α is monotonically increasing on $[a,b]$, $f\in\mathscr{R}(\alpha)$. Define β and g on $[A,B]$ by*

$$\beta(y)=\alpha(\phi(y)),\quad g(y)=f(\phi(y)).$$

Then $g\in\mathscr{R}(\beta)$, and

$$\int_A^B g\,\mathrm{d}\beta=\int_a^b f\,\mathrm{d}\alpha. \tag{19.2}$$

Proof. To each partition $P=\{x_0,\dots,x_n\}$ of $[a,b]$ corresponds a partition $Q=\{y_0,\dots,y_n\}$ of $[A,B]$, where

$$x_i=\phi(y_i)\quad(i=1,\dots,n).$$

All partitions of $[A,B]$ are obtained in this way. Since the values taken by f on $[x_{i-i},x_i]$ are exactly the same as those taken by g on $[y_{i-i},y_i]$, we see that

$$\begin{aligned}U(g,\beta;Q)&=U(f,\alpha;P),\\ L(g,\beta;Q)&=L(f,\alpha;P).\end{aligned} \tag{1}$$

Since $f\in\mathscr{R}(\alpha)$, P can be chosen so that both $U(f,\alpha;P)$ and $L(f,\alpha;P)$ are close to $\int f\,\mathrm{d}\alpha$. Hence (1), combined with 19.4, shows that $g\in\mathscr{R}_\beta[A,B]$ and

$$\int_A^B g\,\mathrm{d}\beta=\int_a^b f\,\mathrm{d}\alpha.$$

□

Note the following special case: Take $\alpha(x)=x$. Then $\beta=\phi$. Assume $\phi'\in\mathscr{R}$. Applying 19.12 to the LHS of

$$\int_A^B g\,\mathrm{d}\beta=\int_a^b f\,\mathrm{d}\alpha,$$

we obtain

$$\int_a^b f(x)\,\mathrm{d}x=\int_A^B f(\phi(y))\,\phi'(y)\,\mathrm{d}y.$$

19.3 Integration and Differentiation

We shall show that integration and differentiation are, in a certain sense, inverse operations.

Theorem 19.14. *Suppose $f \in \mathscr{R}(\alpha)$. For $a \le x \le b$, let the* cumulative function *be*

$$F(x) = \int_a^x f(t)\,\mathrm{d}t\,.$$

Then F is continuous on $[a,b]$; furthermore, if f is continuous at $x_0 \in [a,b]$, then F is differentiable at x_0, and

$$F'(x_0) = f(x_0).$$

Proof. Suppose $f \in \mathscr{R}(\alpha)$. Since f is bounded, let $|f(t)| \le M$ for $t \in [a,b]$. If $a \le x < y \le b$, then

$$\begin{aligned} |F(y) - F(x)| &= \left| \int_a^y f(t)\,\mathrm{d}t - \int_a^x f(t)\,\mathrm{d}t \right| \\ &= \left| \int_x^y f(t)\,\mathrm{d}t \right| \\ &\le \int_x^y |f(t)|\,\mathrm{d}t \\ &\le M(y-x). \end{aligned}$$

Hence F is Lipschitz continuous, so F is uniformly continuous on $[a,b]$.

Now suppose f is continuous at x_0. Fix $\varepsilon > 0$, choose $\delta > 0$ such that for $a \le t \le b$,

$$|t - x_0| < \delta \implies |f(t) - f(x_0)| < \varepsilon.$$

Hence, if s,t are such that

$$x_0 - \delta < s \le x_0 \le t < x_0 + \delta \quad \text{and} \quad a \le x < t \le b,$$

we have, by 19.10(v),

$$\begin{aligned}\left|\frac{F(t)-F(s)}{t-s}-f(x_0)\right| &= \left|\frac{\int_a^s f(u)\,\mathrm{d}u-\int_a^s f(u)\,\mathrm{d}u}{t-s}-f(x_0)\right| \\ &= \left|\frac{1}{t-s}\int_s^t (f(u)-f(x_0))\,\mathrm{d}u\right| \\ &= \frac{1}{t-s}\left|\int_s^t (f(u)-f(x_0))\,\mathrm{d}u\right| \\ &\le \frac{1}{t-s}\int_s^t |f(u)-f(x_0)|\,\mathrm{d}u \\ &< \frac{1}{t-s}\varepsilon(t-s)=\varepsilon\end{aligned}$$

so it follows that $F'(x_0)=f(x_0)$. □

Theorem 19.15 (Fundamental theorem of calculus). *Suppose $f\in\mathscr{R}(\alpha)$, and there exists a differentiable function F on $[a,b]$ such that $F'=f$. Then*

$$\int_a^b f(x)\,\mathrm{d}x = F(b)-F(a). \tag{19.3}$$

Proof. Let $\varepsilon>0$ be given. Choose a partition $P=\{x_0,\dots,x_n\}$ of $[a,b]$ such that $U(f;P)-L(f;P)<\varepsilon$. By the mean value theorem, there exist $t_i\in[x_{i-1},x_i]$ such that

$$\begin{aligned}F(x_i)-F(x_{i-1}) &= F'(t_i)\Delta x_i \\ &= f(t_i)\Delta x_i.\end{aligned}$$

Thus

$$\sum_{i=1}^n f(t_i)\Delta x_i = F(b)-F(a).$$

Then by 19.5,

$$\left|F(b)-F(a)-\int_a^b f(x)\,\mathrm{d}x\right| = \left|\sum_{i=1}^n f(t_i)\Delta x_i-\int_a^b f(x)\,\mathrm{d}x\right|<\varepsilon.$$

Since this holds for all $\varepsilon>0$, the proof is complete. □

Lemma 19.16 (Integration by parts). *Suppose F and G are differentiable on $[a,b]$, $F'=f\in\mathscr{R}$ and $G'=g\in\mathscr{R}$. Then*

$$\int_a^b F(x)g(x)\,\mathrm{d}x = F(b)G(b)-F(a)G(a)-\int_a^b f(x)G(x)\,\mathrm{d}x. \tag{19.4}$$

Proof. Let $H(x) = F(x)G(x)$. Then apply the fundamental theorem of calculus to H and its derivative. □

19.4 Integration of Vector-valued Functions

Let $f_1,\dots,f_k\colon [a,b]\to\mathbb{R}$, and let $\mathbf{f}=(f_1,\dots,f_k)$ where $\mathbf{f}\colon [a,b]\to\mathbb{R}^k$. We say that $\mathbf{f}\in\mathscr{R}(\alpha)$ if $f_1,\dots,f_k\in\mathscr{R}(\alpha)$. If this is the case, we define

$$\int_a^b \mathbf{f}\,d\alpha := \left(\int_a^b f_1\,d\alpha,\dots,\int_a^b f_k\,d\alpha\right).$$

In other words, we "integrate componentwise", so that $\int \mathbf{f}\,d\alpha$ is the point in $\mathbb{R}^k$ whose i-th coordinate is $\int f_i\,d\alpha$.

It is clear that parts (a), (c), and (e) of Theorem 6.12 are valid for these vector-valued integrals; we simply apply the earlier results to each coordinate. The same is true of Theorems 6.17, 6.20, and 6.21. To illustrate, we state the analogue of the fundamental theorem of calculus.

Theorem 19.17. *If $\mathbf{f},\mathbf{F}\colon [a,b]\to\mathbb{R}^k$, $\mathbf{f}\in\mathscr{R}(\alpha)$, and $\mathbf{F}'=\mathbf{f}$. Then*

$$\int_a^b \mathbf{f}(t)\,dt = \mathbf{F}(b)-\mathbf{F}(a). \tag{19.5}$$

The analogue of Theorem 6.13(b) offers some new features, however, at least in its proof.

Lemma 19.18 (Triangle inequality). *Let $\mathbf{f}:[a,b]\to\mathbb{R}^k$, $\mathbf{f}\in\mathscr{R}(\alpha)$ where α is monotonically increasing on $[a,b]$. Then $|\mathbf{f}|\in\mathscr{R}(\alpha)$, and*

$$\left\|\int_a^b \mathbf{f}\,d\alpha\right\| \le \int_a^b \|\mathbf{f}\|\,d\alpha.$$

Proof. If $f_1,\dots,f_k$ are the components of $\mathbf{f}$, then

$$\|\mathbf{f}\| = \left({f_1}^2+\cdots+{f_k}^2\right)^{1/2}.$$

By 19.9, each of the functions ${f_i}^2\in\mathscr{R}(\alpha)$, so their sum ${f_1}^2+\cdots+{f_k}^2\in\mathscr{R}(\alpha)$.

Since x^2 is a continuous function of x, Theorem 4.17 shows that the square-root function is continuous on $[0,M]$, for every real M. If we apply Theorem 6.11 once more, (41) shows that $\|\mathbf{f}\|\in\mathscr{R}(\alpha)$.

Let $\mathbf{y}=(y_1,\dots,y_k)$, where $y_i=\int f_i\,d\alpha$. Then we have $\mathbf{y}=\int \mathbf{f}\,d\alpha$, and

$$\|\mathbf{y}\|^2 = \sum_{i=1}^k {y_i}^2 = \sum_{i=1}^k \left(y_i\int f_i\,d\alpha\right) = \int\left(\sum_{i=1}^k y_i f_i\right)d\alpha.$$

By the Cauchy–Schwarz inequality,

$$\sum_{i=1}^{k} y_i f_i(t) \leq \|\mathbf{y}\| \|\mathbf{f}(t)\| \quad (a \leq t \leq b);$$

hence Theorem 6.12(b) implies

$$\|\mathbf{f}\|^2 \leq \|\mathbf{y}\| \int \|\mathbf{f}\| \, d\alpha .$$

If $\mathbf{y} = \mathbf{0}$, (40) is trivial. If $\mathbf{y} \neq \mathbf{0}$, division of (43) by $\|\mathbf{y}\|$ gives (40). □

to do

19.5 Rectifiable Curves

Definition 19.19 (Curve). A ***curve*** in a set X is a continuous mapping $\gamma\colon [a,b] \to X$. If γ is bijective, γ is called an *arc*. If $\gamma(a) = \gamma(b)$, γ is said to be a *closed curve*.

The case $X = \mathbb{R}^2$ or $X = \mathbb{C}$ (i.e., the case of plane curves) is of considerable importance in the study of analytic functions of a complex variable.

Remark. Note that we define a curve to be a mapping, not a point set. Of course, with each curve γ in $\mathbb{R}^k$ there is associated a subset of $\mathbb{R}^k$, namely the range of γ, but different curves may have the same range.

For each partition $P = \{x_0, \dots, x_n\}$ of $[a,b]$ and each curve γ on $[a,b]$, define

$$\Lambda(\gamma; P) := \sum_{i=1}^{n} |\gamma(x_i) - \gamma(x_{i-1})|.$$

The i-th term in this sum is the distance (in $\mathbb{R}^k$) between the points $\gamma(x_{i-1})$ and $\gamma(x_i)$, Hence $\Lambda(\gamma; P)$ is the length of a polygonal path with vertices at $\gamma(x_0), \gamma(x_1), \dots, \gamma(x_n)$, in this order. As our partition becomes finer and finer, this polygon approaches the range of γ more and more closely.

insert figure

Definition 19.20. The ***total variation*** (or *length*) of γ is

$$\Lambda(\gamma) := \sup_{P \in \mathscr{P}[a,b]} \Lambda(\gamma; P).$$

We say γ is ***rectifiable*** if $\Lambda(\gamma) < \infty$.

The next result gives a formula for calculating the length of a rectifiable curve that is continuously differentiable.

Proposition 19.21. *If γ is a continuously differentiable curve on $[a,b]$, then γ is rectifiable, and*

$$\Lambda(\gamma) = \int_a^b |\gamma'(t)|\,\mathrm{d}t\,. \tag{19.6}$$

Proof. If $a \le x_{i-1} < x_i \le b$, then

$$|\gamma(x_i) - \gamma(x_{i-1})| = \left| \int_{x_{i-1}}^{x_i} \gamma'(t)\,\mathrm{d}t \right| \le \int_{x_{i-1}}^{x_i} |\gamma'(t)|\,\mathrm{d}t\,.$$

Hence, for every partition P of $[a,b]$, taking the sum on both sides gives

$$\Lambda(\gamma;P) \leq \int_a^b |\gamma'(t)|\,\mathrm{d}t$$

and taking sup gives

$$\Lambda(\gamma) \leq \int_a^b |\gamma'(t)|\,\mathrm{d}t\,.$$

We now prove the opposite inequality. Since γ' is (continuous and thus) uniformly continuous on $[a,b]$, fix $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|s-t| < \delta \implies |\gamma'(s) - \gamma'(t)| < \varepsilon.$$

Let $P = \{x_0, \ldots, x_n\}$ be a partition of $[a,b]$, with $\Delta x_i < \delta$ for all i. If $t \in [x_{i-1}, x_i]$, it follows that

$$|\gamma'(t)| \leq |\gamma'(x_i)| + \varepsilon.$$

Hence

$$\begin{aligned}
\int_{x_{i-1}}^{x_i} |\gamma'(t)|\,\mathrm{d}t &\leq |\gamma'(x_i)|\Delta x_i + \varepsilon\Delta x_i \\
&= \left| \int_{x_{i-1}}^{x_i} \big(\gamma'(t) + \gamma'(x_i) - \gamma'(t)\big)\,\mathrm{d}t \right| + \varepsilon\Delta x_i \\
&\leq \left| \int_{x_{i-1}}^{x_i} \gamma'(t)\,\mathrm{d}t \right| + \left| \int_{x_{i-1}}^{x_i} \big(\gamma'(x_i) - \gamma'(t)\big)\,\mathrm{d}t \right| + \varepsilon\Delta x_i \\
&\leq |\gamma(x_i) - \gamma(x_{i-1})| + 2\varepsilon\Delta x_i.
\end{aligned}$$

If we add these inequalities, we obtain

$$\begin{aligned}
\int_a^b |\gamma'(t)|\,\mathrm{d}t &\leq \Lambda(\gamma;P) + 2\varepsilon(b-a) \\
&\leq \Lambda(\gamma) + 2\varepsilon(b-a).
\end{aligned}$$

Since ε was arbitrary, we must have

$$\int_a^b |\gamma'(t)| \leq \Lambda(\gamma).$$

This completes the proof. □

Exercises

20 Sequences and Series of Functions

Suppose $f_n : E \subset X \to Y$ is a sequence of functions. In some cases, we shall restrict ourselves to complex-valued functions (take $Y = \mathbb{C}$).

20.1 Pointwise Convergence

A natural extension of convergence of sequences of numbers to sequences of functions is to fix a point $x \in E$, and consider the behaviour of the sequence $(f_n(x))$.

Definition 20.1 (Pointwise convergence)**.** Suppose (f_n) is a sequence of functions, and $(f_n(x))$ converges for every $x \in E$. We say (f_n) ***converges pointwise*** to f on E, denoted by $f_n \to f$, if

$$f(x) = \lim_{n\to\infty} f_n(x) \quad (\forall x \in E).$$

That is, for all $x \in E$,

$$\forall \varepsilon > 0, \quad \exists N \in \mathbb{N}, \quad \forall n \geq N, \quad d\left(f_n(x) - f(x)\right) < \varepsilon.$$

f is called the *limit* (or *limit function*) of (f_n).

Similarly, if $\sum f_n(x)$ converges for every $x \in E$, and if we define

$$f(x) = \sum_{n=1}^{\infty} f_n(x) \quad (\forall x \in E)$$

the function f is called the *sum of the series* $\sum f_n$.

Example. The sequence of functions $f_n(x) = \frac{x}{n}$ converges pointwise to the zero function $f(x) = 0$.

The main problem which arises is to determine whether important properties of functions are preserved by pointwise convergence. For instance, if f_n are continuous, or differentiable, or integrable, is the same true of the limit function? What are the relations between f_n' and f', say,

or between $\int f_n$ and $\int f$?

Example (Continuity). For $0 < x < 1$, the sequence of functions $f_n(x) = x^n$ converges pointwise to the function

$$f(x) = \begin{cases} 1 & (x = 1) \\ 0 & (0 \le x < 1) \end{cases}$$

Evidently f_n are continuous, but f is discontinuous. Hence

$$\lim_{x \to x_0} \lim_{n \to \infty} f_n(x) \neq \lim_{n \to \infty} \lim_{x \to x_0} f_n(x).$$

Example (Differentiability). For $x \in \mathbb{R}$, let

$$f_n(x) = \frac{\sin nx}{\sqrt{n}} \quad (n = 1, 2, \dots)$$

so

$$f(x) = \lim_{n \to \infty} f_n(x) = 0.$$

Then $f'(x) = 0$, and

$$f_n'(x) = \sqrt{n} \cos nx,$$

so (f_n') does not converge to f'.

This shows that the limit of the derivative does not equal the derivative of the limit.

Example (Integrability). Let

$$f_n(x) = \chi_{[n,n+1]}(x),$$

Then $\displaystyle\int_{\mathbb{R}} f_n(x)\,\mathrm{d}x = 1$, so

$$\lim_{n \to \infty} \int_{\mathbb{R}} f_n(x)\,\mathrm{d}x = 1.$$

However

$$\int_{\mathbb{R}} \lim_{n \to \infty} f_n(x)\,\mathrm{d}x = \int 0\,\mathrm{d}x = 0.$$

This shows that the limit of the integral does not equal the integral of the limit. Thus we may not switch the order of limits.

Pointwise convergence does not preserve many nice properties of functions. Hence, we need a stronger notion of convergence for sequences and series of functions.

20.2 Uniform Convergence

Definition 20.2 (Uniform convergence)**.** We say (f_n) ***converges uniformly*** to f on E, denoted by $f_n \rightrightarrows f$, if

$$\forall \varepsilon > 0, \quad \exists N \in \mathbb{N}, \quad \forall x \in E, \quad \forall n \geq N, \quad d\left(f_n(x) - f(x)\right) < \varepsilon.$$

Similarly, a series of functions $\sum f_n(x)$ converges uniformly on E if the sequence of partial sums (s_n) defined by

$$s_n(x) = \sum_{k=1}^{n} f_k(x)$$

converges uniformly on E.

Intuitively, uniform convergence can be visualised as the sequence of functions (f_n) eventually contained in an ε-tube around f, for sufficiently large n.

insert figure

Remark. Uniform convergence is stronger than pointwise convergence, since N is uniform (or "fixed") for all $x \in E$; for pointwise convergence, the choice of N is determined by x.

Uniform convergence implies pointwise convergence, but not the other way around.

Example. Consider the sequence of functions $f_n(x) = x^n$ defined on $(0,1)$. Then $f_n \to 0$. But $f_n \not\rightrightarrows 0$.

Proof. □

For the rest of this chapter, unless stated otherwise, we shall restrict our focus to sequences of complex-valued functions defined on $E \subset X$.

We give a name to sequences of functions that satisfy the Cauchy criterion.

Definition 20.3. We say (f_n) is *uniformly Cauchy* if

$$\forall \varepsilon > 0, \quad \exists N \in \mathbb{N}, \quad \forall x \in E, \quad \forall n, m \geq N, \quad |f_n(x) - f_m(x)| < \varepsilon.$$

Lemma 20.4 (Cauchy criterion)**.** *$f_n \rightrightarrows f$ on E if and only if (f_n) is uniformly Cauchy.*

Proof.

$\boxed{\Longrightarrow}$ Suppose $f_n \rightrightarrows f$ on E. Let $\varepsilon > 0$ be given. There exists $N \in \mathbb{N}$ such that for all $x \in E$, for all $n \geq N$,

$$|f_n(x) - f(x)| < \frac{\varepsilon}{2}.$$

Then for all $n, m \geq N$,

$$\begin{aligned} |f_n(x) - f_m(x)| &= |(f_n(x) - f(x)) + (f(x) - f_m(x))| \\ &\leq |f_n(x) - f(x)| + |f_m(x) - f(x)| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

$\boxed{\Longleftarrow}$ Suppose (f_n) is uniformly Cauchy.

For every $x \in E$, the sequence $(f_n(x))$ is a Cauchy sequence and thus converges to a limit $f(x)$. Hence by definition, $f_n \to f$ on E. We are left to prove that the convergence is uniform.

Let $\varepsilon > 0$ be given. There exists $N \in \mathbb{N}$ such that for all $n, m \geq N$ and for all $x \in E$,

$$|f_n(x) - f_m(x)| < \varepsilon.$$

Fix n, and let $m \to \infty$. Since $\lim_{m\to\infty} f_m(x) = f(x)$, thus for all $n \geq N$ and for all $x \in E$,

$$|f_n(x) - f(x)| < \varepsilon,$$

which completes the proof. □

Definition 20.5. Let $f \in \mathscr{C}(X)$. We define the ***suprenum norm*** of f as

$$\|f\| := \sup_{x\in X} |f(x)|.$$

We check that $\|f\|$ gives a norm on $\mathscr{C}(X)$:

(i) $|f(x)| \geq 0$ for all $x \in X$, so $\|f\| \geq 0$. It is clear that $\|f\| = 0$ if and only if $f(x) = 0$ for every $x \in X$, that is, only if $f = 0$.

(ii) For all $\lambda \in \mathbb{C}$,
$$\|\lambda f\| = \sup_{x\in X} |\lambda f(x)| = |\lambda| \sup_{x\in X} |f(x)| = |\lambda| \|f\|.$$

(iii) If $h = f + g$, then for all $x \in X$,

$$|h(x)| \leq |f(x)| + |g(x)| \leq \|f\| + \|g\|.$$

Hence taking sup on the left gives $\|f + g\| \leq \|f\| + \|g\|$.

The metric induced by the suprenum norm is

$$d(f, g) = \|f - g\|.$$

With this metric, $\mathscr{C}(X)$ is a metric space.

The following result provides another equivalent way to determine uniform convergence.

Lemma 20.6. *$f_n \rightrightarrows f$ on E if and only if $f_n \to f$ on E with respect to the metric of $\mathscr{C}(E)$.*

Proof.

$$\begin{aligned} f_n \to f &\iff \lim_{n\to\infty} \|f_n - f\| = 0 \\ &\iff \lim_{n\to\infty} \left(\sup_{x\in E} |f_n(x) - f(x)| \right) = 0 \\ &\iff \forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, \sup_{x\in E} |f_n(x) - f(x)| < \varepsilon \\ &\iff \forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, \forall x \in E, |f_n(x) - f(x)| < \varepsilon \end{aligned}$$

which precisely means that $f_n \rightrightarrows f$ on E, by definition.

Note that the $\boxed{\Longleftarrow}$ direction of the last step is tricky, since the limit can equal ε, so we take $\frac{\varepsilon}{2}$ instead. □

For series, there is a very convenient test for uniform convergence, due to Weierstrass.

Lemma 20.7 (Weierstrass M-test)**.** *Suppose (f_n) is a sequence of complex-valued functions defined on E, and*

$$|f_n(x)| \leq M_n \quad (n = 1, 2, \ldots, x \in E)$$

If $\sum M_n$ converges, then $\sum f_n$ converges uniformly on E.

Proof. Suppose $\sum M_n$ converges. Let $\varepsilon > 0$ be given, the partial sums of $\sum M_n$ form a Cauchy sequence, so there exists $N \in \mathbb{N}$ such that for all $n \geq m \geq N$,

$$\sum_{k=m}^{n} M_k < \varepsilon.$$

Considering the partial sums of the series of functions,

$$\left| \sum_{k=m}^{n} f_k(x) \right| \leq \sum_{k=m}^{n} |f_k(x)| \leq \sum_{k=m}^{n} M_k < \varepsilon.$$

By the Cauchy criterion (20.4), we are done. □

Example.

- The series $\sum_{n=1}^{\infty} \frac{\sin nx}{n^2}$ converges uniformly on $\mathbb{R}$. (Note: this is a Fourier series, we'll see more of these later). That is because

$$\left|\frac{\sin nx}{n^2}\right| \leq \frac{1}{n^2} \quad \text{and} \quad \sum_{n=1}^{\infty} \frac{1}{n^2} \text{ converges.}$$

- The series $\sum_{n=0}^{\infty} \frac{x^n}{n!}$ converges uniformly on any bounded interval. For example take the interval $[-r,r] \subset \mathbb{R}$,

$$\left|\frac{x^n}{n!}\right| \leq \frac{r^n}{n!} \quad \text{and} \quad \sum_{n=1}^{\infty} \frac{r^n}{n!} \text{ converges by the ratio test.}$$

20.3 Properties of Uniform Convergence

We now consider properties preserved by uniform convergence.

Uniform Convergence and Continuity

We prove a more general result.

Proposition 20.8. *Suppose $f_n \rightrightarrows f$ on E. Let $x \in X$ be a limit point of E, and suppose that*

$$\lim_{t\to x} f_n(t) = A_n \quad (n = 1,2,\dots).$$

Then (A_n) converges, and $\lim_{t\to x} f(t) = \lim_{n\to\infty} A_n$.

In other words, the conclusion is that

$$\lim_{t\to x}\lim_{n\to\infty} f_n(t) = \lim_{n\to\infty}\lim_{t\to x} f_n(t).$$

Proof.

1. We first show that (A_n) converges. Since (f_n) uniformly converges on E, by the Cauchy criterion (20.4), fix $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $n, m \geq N$, $t \in E$,

   $$|f_n(t) - f_m(t)| < \varepsilon.$$

   Letting $t \to x$, since $\lim_{t\to x} f_n(t) = A_n$, we have that for all $n, m \geq N$,

   $$|A_n - A_m| < \varepsilon.$$

   Thus (A_n) is a Cauchy sequence and therefore converges, say to A.

2. Next we will show that $\lim_{t\to x} f(t) = A$. Write

   $$|f(t) - A| \leq |f(t) - f_n(t)| + |f_n(t) - A_n| + |A_n - A|. \tag{1}$$

   By the uniform convergence of (f_n), there exists $N_1 \in \mathbb{N}$ such that for all $n \geq N_1$,

   $$|f(t) - f_n(t)| < \frac{\varepsilon}{3} \quad (t \in E).$$

   By the convergence of (A_n), there exists $N_2 \in \mathbb{N}$ such that for all $n \geq N_2$,

   $$|A_n - A| < \frac{\varepsilon}{3}.$$

Choose $N = \max\{N_1, N_2\}$ such that the above two inequalities hold simultaneously. Then for this n, since $\lim_{t\to x} f_n(t) = A_n$, we choose an open ball B of x such that if $t \in B \cap E$, $t \neq x$, then

$$|f_n(t) - A_n| < \frac{\varepsilon}{3}.$$

Substituting the above inequalities into (1) gives

$$|f(t) - A| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.$$

provided $t \in B \cap E$, $t \neq x$. This is equivalent to $\lim_{t\to x} f(t) = A$.

□

An important corollary is that uniform convergence preserves continuity.

Corollary 20.9. *Suppose (f_n) are continuous on E, and $f_n \rightrightarrows f$ on E. Then f is continuous on E.*

Proof. By continuity of f_n,

$$\lim_{t\to x} f_n(t) = f_n(x).$$

Then

$$\lim_{t\to x} f(t) = \lim_{t\to x}\left(\lim_{n\to\infty} f_n(t)\right) = \lim_{n\to\infty}\left(\lim_{t\to x} f_n(t)\right) = \lim_{n\to\infty} f_n(x) = f(x),$$

which precisely means that f is continuous on E. □

Remark. The converse is not true; for instance, the sequence of functions $f_n : (0,1) \to \mathbb{R}$ defined by $f_n(x) = x^n$ converges to the zero function, which is continuous, but the convergence is not uniform.

Let us see that we can have extra conditions such that the converse of the previous result is true.

Proposition 20.10 (Dini's theorem)**.** *Suppose K is compact, and (f_n) is a sequence of continuous functions on K, $f_n \to f$ on K, and (f_n) is monotonically decreasing:*

$$f_n(x) \geq f_{n+1}(x) \quad (n = 1, 2, \ldots).$$

Then $f_n \rightrightarrows f$ on K.

Remark. The compactness in the hypotheses is necessary; for instance, on $(0,1)$ define $f_n(x) = \frac{1}{nx+1}$. Then $f_n(x) \to 0$ monotonically in $(0,1)$, but the convergence is not uniform.

Proof. Let $g_n = f_n - f$. Then g_n is continuous, $g_n \to 0$, and $g_n \geq g_{n+1} \geq 0$. We have to prove that $g_n \rightrightarrows 0$ on K.

Let $\varepsilon > 0$ be given. For $n = 1, 2, \ldots$, let

$$K_n = \{x \in K \mid g_n(x) \geq \varepsilon\}.$$

Since g_n is continuous, and $\{g_n(x) \mid g_n(x) \geq \varepsilon\}$ is closed, by 17.16, its pre-image K_n is closed. Since K_n is a closed subset of a compact set K, by 15.32, K_n is compact.

Since $g_n \geq g_{n+1}$, we have $K_n \supset K_{n+1}$. Fix $x \in K$. Since $g_n(x) \to 0$, we see that $x \notin K_n$ if n is sufficiently large. Thus $x \notin \bigcap_{n=1}^{\infty} K_n$. In other words, $\bigcap_{n=1}^{\infty} K_n = \emptyset$. Hence $K_N = \emptyset$ for some N (by the converse of Cantor's intersection theorem). It follows that for all $x \in K$ and for all $n \geq N$,

$$0 \leq g_n(x) < \varepsilon.$$

Therefore $g_n \rightrightarrows 0$ on K, as desired. □

Lemma 20.11. *$\mathscr{C}(X)$ is a complete metric space.*

Proof. Let (f_n) be a Cauchy sequence in $\mathscr{C}(X)$. Then fix $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $n, m \geq N$,

$$\|f_n - f_m\| < \varepsilon.$$

By the Cauchy criterion (20.4), $f_n \rightrightarrows f$ for some $f \colon X \to \mathbb{C}$. We now need to show that $f \in \mathscr{C}(X)$; that is, f is continuous and bounded.

- f is continuous by 20.9.
- f is bounded, since there is an n such that $|f(x) - f_n(x)| < 1$ for all $x \in X$, and f_n is bounded.

Hence $f \in \mathscr{C}(X)$, and since $f_n \rightrightarrows f$ on X, we have $\|f - f_n\| \to 0$ as $n \to \infty$. □

Uniform Convergence and Integration

The next result states that the limit and integral can be interchanged.

Proposition 20.12. *Suppose $f_n \rightrightarrows f$ on $[a,b]$, $f_n \in \mathscr{R}(\alpha)$ and $\alpha \nearrow$. Then $f \in \mathscr{R}(\alpha)$, and*

$$\lim_{n\to\infty} \int_a^b f_n \,\mathrm{d}\alpha = \int_a^b f \,\mathrm{d}\alpha. \tag{20.1}$$

Proof. It suffices to prove this for real-valued f_n. Let

$$\varepsilon_n = \sup_{x\in[a,b]} |f_n(x) - f(x)|.$$

Then $|f_n - f| \le \varepsilon$, so

$$f_n - \varepsilon_n \le f \le f_n + \varepsilon_n,$$

so that the upper and lower integrals of f satisfy

$$\int_a^b (f_n - \varepsilon_n)\,\mathrm{d}\alpha \le \underline{\int_a^b} f \,\mathrm{d}\alpha \le \overline{\int_a^b} f \,\mathrm{d}\alpha \le \int_a^b (f_n + \varepsilon_n)\,\mathrm{d}\alpha.$$

Hence

$$0 \le \overline{\int_a^b} f \,\mathrm{d}\alpha - \underline{\int_a^b} f \,\mathrm{d}\alpha \le 2\varepsilon_n[\alpha(b) - \alpha(a)].$$

Since $f_n \rightrightarrows f$, we see that $\varepsilon_n \to 0$ as $n \to \infty$, the upper and lower integrals of f are equal. Hence $f \in \mathscr{R}(\alpha)$.

We have

$$\begin{aligned}
\left| \int_a^b f_n \,\mathrm{d}\alpha - \int_a^b f \,\mathrm{d}\alpha \right| &= \left| \int_a^b f_n - f \,\mathrm{d}\alpha \right| \\
&\le \int_a^b |f_n - f| \,\mathrm{d}\alpha \\
&\le [\alpha(b) - \alpha(a)] \sup_{x\in[a,b]} |f_n(x) - f(x)| \\
&= \varepsilon_n[\alpha(b) - \alpha(a)].
\end{aligned}$$

This implies

$$\lim_{n\to\infty} \int_a^b f_n \,\mathrm{d}\alpha = \int_a^b f \,\mathrm{d}\alpha.$$

□

Corollary 20.13. *Suppose $f_n \in \mathscr{R}(\alpha)$ and*

$$f(x) = \sum_{n=1}^{\infty} f_n(x)$$

converges uniformly on $[a,b]$. Then

$$\int_a^b f \, \mathrm{d}\alpha = \sum_{n=1}^{\infty} \int_a^b f_n \, \mathrm{d}\alpha .$$

In other words, we can swap the integral and sum, such that the series may be integrated term by term.

Proof. Consider the sequence of partial sums

$$f_n(x) = \sum_{k=1}^{n} f_k(x) \quad (n = 1, 2, \dots).$$

It follows $f_n \in \mathscr{R}(\alpha)$ and $f_n \rightrightarrows f$. Apply above theorem to (f_n) and the conclusion follows. □

Example. Let us show how to integrate a Fourier series:

$$\int_0^x \sum_{n=1}^{\infty} \frac{\cos nt}{n^2} \, \mathrm{d}t = \sum_{n=1}^{\infty} \int_0^x \frac{\cos nt}{n^2} \, \mathrm{d}t = \sum_{n=1}^{\infty} \frac{\sin nx}{n^3}.$$

Uniform Convergence and Differentiation

The next result shows that the process of limit and differentiation can be interchanged.

Proposition 20.14. *Suppose (f_n) are differentiable on $[a,b]$, and $(f_n(x_0))$ converges for some $x_0 \in [a,b]$. If f_n' converges uniformly on $[a,b]$, then there exists a differentiable f such that $f_n \rightrightarrows f$ on $[a,b]$, and*

$$f'(x) = \lim_{n\to\infty} f_n'(x) \quad (a \le x \le b). \tag{20.2}$$

Proof. We first show that (f_n) converges uniformly on $[a,b]$, then show that the limit f is differentiable, and finally show that (20.2) holds.

1. Let $\varepsilon > 0$ be given. Since $(f_n(x_0))$ converges, $(f_n(x_0))$ is a Cauchy sequence; choose $N \in \mathbb{N}$ such that for all $n,m \ge N$,

   $$|f_n(x_0) - f_m(x_0)| < \frac{\varepsilon}{2}.$$

   Since (f_n') converges uniformly on $[a,b]$, by 20.4, (f_n') is uniformly Cauchy. Thus

   $$\left|f_n'(x) - f_m'(x)\right| < \frac{\varepsilon}{2(b-a)} \quad (a \le x \le b).$$

   We now apply the mean value theorem (18.15) to the function $f_n - f_m$: for any $x,t \in [a,b]$, if $n,m \ge N$, then

   $$|(f_n(x) - f_m(x)) - (f_n(t) - f_m(t))| < \frac{\varepsilon}{2(b-a)}|x-t| \le \frac{\varepsilon}{2} \tag{1}$$

   Finally, by the triangle inequality,

   $$\begin{aligned} |f_n(x) - f_m(x)| &\le |f_n(x) - f_m(x) - f_n(x_0) + f_m(x_0)| + |f_n(x_0) - f_m(x_0)| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

   This holds true for all $x \in [a,b]$. Hence by 20.4, (f_n) converges uniformly on $[a,b]$.

2. Let

   $$f(x) = \lim_{n\to\infty} f_n(x) \quad (a \le x \le b).$$

   Fix a point $x \in [a,b]$, and let

   $$\phi_n(t) = \frac{f_n(t) - f_n(x)}{t-x}, \quad \phi(t) = \frac{f(t) - f(x)}{t-x} \quad (a \le t \le b, t \ne x).$$

Idea. To show that f is differentiable, we need to show that $\lim_{t\to x}\phi(t)$ exists.

Note that since f_n are differentiable, we have

$$\lim_{t\to x}\phi_n(t) = f_n'(x) \quad (n = 1,2,\dots).$$

By (1), for all $n,m \geq N$,

$$\begin{aligned}|\phi_n(t)-\phi_m(t)| &= \frac{1}{|t-x|}\,|(f_n(t)-f_n(x))-(f_m(t)-f_m(x))| \\ &= \frac{1}{|x-t|}\,|(f_n(x)-f_m(x))-(f_n(t)-f_m(t))| \\ &< \frac{1}{\cancel{|x-t|}}\cdot\frac{\varepsilon}{2(b-a)}\cancel{|x-t|} = \frac{\varepsilon}{2(b-a)},\end{aligned}$$

so (ϕ_n) converges uniformly, for $t \neq x$. Since (f_n) converges to f, we conclude that

$$\lim_{n\to\infty}\phi_n(t) = \lim_{n\to\infty}\frac{f_n(t)-f_n(x)}{t-x} = \frac{f(t)-f(x)}{t-x} = \phi(t)$$

uniformly for $a \leq t \leq b$, $t \neq x$.

Applying 20.8 to (ϕ_n), we obtain

$$\lim_{t\to x}\phi(t) = \lim_{t\to x}\lim_{n\to\infty}\phi_n(t) = \lim_{n\to\infty}\lim_{t\to x}\phi_n(t) = \lim_{n\to\infty}f_n'(x),$$

which is precisely (20.2).

□

Example (Weierstrass function)**.** We will now construct a continuous nowhere differentiable function on $\mathbb{R}$. Define

$$\phi(x) = |x| \quad (-1 \leq x \leq 1).$$

We extend the definition of $\phi(x)$ to all of $\mathbb{R}$ by making ϕ 2-periodic: $\phi(x) = \phi(x+2)$. Then $\phi : \mathbb{R} \to \mathbb{R}$ is continuous as $|\phi(x)-\phi(y)| \leq |x-y|$ (not hard to prove).
Let the *Weierstrass function* be defined as

$$f(x) = \sum_{n=0}^{\infty}\left(\frac{3}{4}\right)^n \phi(4^n x).$$

Claim. The Weierstrass function is continuous and nowhere differentiable on $\mathbb{R}$.

- Since $\sum\left(\frac{3}{4}\right)^n$ converges, and $|\phi(x)| \leq 1$ for all $x \in \mathbb{R}$, by the Weierstrass M-test,

$f(x)$ converges uniformly and hence is continuous.

- Fix $x \in \mathbb{R}$ and $m \in \mathbb{Z}^+$, and define

$$\delta_m = \pm\frac{1}{2}\cdot 4^{-m},$$

where the sign is chosen in such a way so that there is no integer between $4^m x$ and $4^m(x+\delta_m)$, which can be done since $4^m|\delta_m| = \frac{1}{2}$. Define

$$\gamma_n = \frac{\phi\left(4^n(x+\delta_m)\right) - \phi(4^n x)}{\delta_m}.$$

If $n > m$, then as $4^n\delta_m$ is an even integer. Then as ϕ is 2-periodic we get that $\gamma_n = 0$.

Furthermore, since there is no integer between $4^m x \pm \frac{1}{2}$ and $4^m x$, we have that

$$\left|\phi\left(4^m x \pm \frac{1}{2}\right) - \phi(4^m x)\right| = \left|\left(4^m x \pm \frac{1}{2}\right) - 4^m x\right| = \frac{1}{2}.$$

Therefore

$$|\gamma_n| = \left|\frac{\phi\left(4^m x \pm \frac{1}{2}\right) - \phi(4^m x)}{\pm\frac{1}{2}\cdot 4^{-m}}\right| = 4^m.$$

Similarly, if $n < m$, since $|\phi(s) - \phi(t)| \le |s-t|$,

$$|\gamma_n| = \left|\frac{\phi\left(4^n x \pm \frac{1}{2}\cdot 4^{n-m}\right) - \phi(4^n x)}{\pm\frac{1}{2}\cdot 4^{-m}}\right| \le \left|\frac{\pm\frac{1}{2}\cdot 4^{n-m}}{\pm\frac{1}{2}\cdot 4^{-m}}\right| = 4^n.$$

Finally,

$$\begin{aligned}
\left|\frac{f(x+\delta_m) - f(x)}{\delta_m}\right| &= \left|\sum_{n=0}^{\infty}\left(\frac{3}{4}\right)^n \frac{\phi(4^n(x+\delta_m)) - \phi(4^n x)}{\delta_m}\right| = \left|\sum_{n=0}^{\infty}\left(\frac{3}{4}\right)^n \gamma_n\right| \\
&= \left|\sum_{n=0}^{m}\left(\frac{3}{4}\right)^n \gamma_n\right| \\
&\ge \left|\frac{3}{4}^m \gamma_m\right| - \left|\sum_{n=0}^{m-1}\left(\frac{3}{4}\right)^n \gamma_n\right| \\
&\ge 3^m - \sum_{n=0}^{m-1} 3^n = 3^m - \frac{3^m - 1}{3-1} = \frac{3^m+1}{2}.
\end{aligned}$$

It is obvious that $\delta_m \to 0$ as $m \to \infty$, but $\frac{3^m+1}{2}$ goes to infinity. Hence f cannot be differentiable at x.

20.4 Equicontinuous Families of Functions

We would like an analogue of Bolzano–Weierstrass; that is, every bounded sequence of functions has a convergent subsequence.

Definition 20.15. Suppose (f_n) is a sequence of functions. We say (f_n) is ***pointwise bounded*** on E if for every $x \in E$, the sequence $(f_n(x))$ is bounded; that is,

$$\forall x \in E, \quad \exists M \in \mathbb{R}, \quad \forall n \in \mathbb{N}, \quad |f_n(x)| \leq M.$$

We say (f_n) is ***uniformly bounded*** on E if

$$\exists M \in \mathbb{R}, \quad \forall x \in E, n \in \mathbb{N}, \quad |f_n(x)| \leq M.$$

Lemma 20.16. *Suppose (f_n) is a pointwise bounded sequence of complex-valued functions on a countable set E. Then (f_n) has a subsequence (f_{n_k}) such that $f_{n_k}(x)$ converges for every $x \in E$.*

Proof. We will use a very common and useful diagonal argument.

Arrange the points of E in a sequence (x_i), where $i = 1, 2, \ldots$.

Since (f_n) is pointwise bounded on E, the sequence $(f_n(x_1))_{n=1}^{\infty}$ is bounded. By the Bolzano–Weierstrass theorem, there exists a subsequence, which we denote by $(f_{1,k})_{k=1}^{\infty}$, such that $\big(f_{1,k}(x_1)\big)_{k=1}^{\infty}$ converges.

Consider the array formed by the sequences $S_1, S_2, \ldots$:

$$\begin{array}{lcccc} S_1: & f_{1,1} & f_{1,2} & f_{1,3} & \cdots \\ S_2: & f_{2,1} & f_{2,2} & f_{2,3} & \cdots \\ S_3: & f_{3,1} & f_{3,2} & f_{3,3} & \cdots \\ \vdots & & & & \end{array}$$

and which have the following properties:

(i) S_n is a subsequence of S_{n-1}, for $n = 2, 3, \ldots$

(ii) $\big(f_{n,k}(x_n)\big)$ converges, as $k \to \infty$ (the boundedness of $(f_n(x_n))$ makes it possible to choose S_n in this way);

(iii) The order in which the functions appear is the same in each sequence; i.e., if one function precedes another in S_1, they are in the same relation in every S_n, until one or the other is deleted. Hence, when going from one row in the above array to the next below, functions may move to the left but never to the right.

We now go down the diagonal of the array; i.e., we consider the sequence

$$S: \quad f_{1,1} \quad f_{2,2} \quad f_{3,3} \quad \cdots$$

By (iii), the sequence S (except possibly its first $n-1$ terms) is a subsequence of S_n, for $n = 1, 2, \ldots$. Hence (ii) implies that $(f_{n,n}(x_i))$ converges, as $n \to \infty$, for every $x_i \in E$. □ to do

Definition 20.17. A family $\mathscr{F}$ of functions $f: E \subset X \to \mathbb{C}$ is ***equicontinuous*** on E if

$$\forall \varepsilon > 0, \quad \exists \delta > 0, \quad \forall x, y \in E,\ f \in \mathscr{F}, \quad d(x,y) < \delta \implies |f(x) - f(y)| < \varepsilon.$$

Proposition 20.18. *Suppose X is a compact metric space, $f_n \in \mathscr{C}(X)$, and (f_n) converges uniformly on X. Then (f_n) is equicontinuous on X.*

Proof. Let $\varepsilon > 0$ be given. Since (f_n) converges uniformly on X, $f_n \to f$ on X with respect to the metric of $\mathscr{C}(X)$. Then

$$\lim_{n\to\infty} \|f_n - f\| = 0,$$

i.e., there exists $N \in \mathbb{N}$ such that for all $n \geq N$,

$$\|f_n - f_N\| < \frac{\varepsilon}{3}.$$

Since continuous functions are uniformly continuous on compact sets, f_n are uniformly continuous on K, so there exists $\delta > 0$ such that

$$d(x,y) < \delta \implies |f_i(x) - f_i(y)| < \frac{\varepsilon}{3}$$

for $i = 1, \ldots, N$. If $n \geq N$ and $d(x,y) < \delta$,

$$\begin{aligned} |f_n(x) - f_n(y)| &\leq |f_n(x) - f_N(x)| + |f_N(x) - f_N(y)| + |f_N(y) - f_n(y)| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon. \end{aligned}$$

In conjunction with (43), this proves the theorem. □

We first need the following lemma.

Lemma 20.19. *A compact metric space X contains a countable dense subset.*

Proof. For each $n \in \mathbb{N}$, there exist finitely many balls of radius $\frac{1}{n}$ that cover X (by compactness of X). That is, for every n, there exist finitely many points $x_{n,1}, \ldots, x_{n,k_n}$ such that

$$X = \bigcup_{i=1}^{k_n} B_{\frac{1}{n}}(x_{n,i}).$$

Claim. $S = \{x_{n,i} \mid i = 1, \ldots, k_n\}$ is a countable dense subset of X.

- Since S is a countable union of finite sets, S is countable.
- For every $x \in X$ and every $\varepsilon > 0$, there exists $n \in \mathbb{N}$ such that $\frac{1}{n} < \varepsilon$ and an $x_{n,i} \in S$ such that
$$x \in B_{\frac{1}{n}}(x_{n,i}) \subset B_\varepsilon(x_{n,i}).$$
Hence $x \in \overline{S}$, so $\overline{S} = X$ and therefore S is dense.

□

We can now prove the very useful Arzelà–Ascoli theorem about existence of convergent subsequences.

Theorem 20.20 (Arzelà–Ascoli theorem). *Suppose X is compact, $f_n \in \mathscr{C}(X)$, and (f_n) is pointwise bounded and equicontinuous on X. Then (f_n) is uniformly bounded on X, and contains a uniformly convergent subsequence.*

Proof. Let us first show that the sequence is uniformly bounded. By equicontinuity, there exists $\delta > 0$ such that
$$B_\delta(x) \subset f_n^{-1}(B_1(f_n(x))) \quad (x \in X).$$
Since X is compact, there exist finitely many points $x_1, \ldots, x_k$ such that
$$X = \bigcup_{j=1}^{k} B_\delta(x_j).$$
Since (f_n) is pointwise bounded, there exist $M_1, \ldots, M_k$ such that
$$|f_n(x_j)| \leq M_j \quad (j = 1, \ldots, k)$$
for all n. Let $M = 1 + \max\{M_1, \ldots, M_k\}$. Now given any $x \in X$, $x \in B_\delta(x_j)$ for some $1 \leq j \leq k$. Therefore, for all n we have $x \in f_n^{-1}\left(B_1(f_n(x_j))\right)$ or in other words
$$|f_n(x) - fn(x_j)| < 1.$$
By reverse triangle inequality,
$$|f_n(x)| < 1 + |fn(x_j)| \leq 1 + M_j \leq M$$
Since x was arbitrary, (f_n) is uniformly bounded.

Next, pick a countable dense set S. By Theorem 7.23, there exists a subsequence (f_{n_j}) that converges pointwise on S. Write $g_j = f_{n_j}$ for simplicity. Note that (g_n) is equicontinuous.

Let $\varepsilon > 0$ be given, then pick $\delta > 0$ such that for all $x \in X$,

$$B_\delta(x) \subset g_n^{-1}\left(B_{\frac{\varepsilon}{3}}(g_n(x))\right).$$

By density of S, every $x \in X$ is in some $B_\delta(y)$ for some $y \in S$, and by compactness of X, there is a finite subset $\{x_1, \ldots, x_k\}$ of S such that

$$X = \bigcup_{j=1}^{k} B_\delta(x_j).$$

Now as there are finitely many points and we know that (g_n) converges pointwise on S, there exists $N \in \mathbb{N}$ such that for all $n, m \geq N$,

$$|g_n(x_j) - g_m(x_j)| < \frac{\varepsilon}{3} \quad (j = 1, \ldots, k).$$

Let $x \in X$ be arbitrary. There is some i such that $x \in B_\delta(x_i)$ and so we have for all $i \in \mathbb{N}$,

$$|g_i(x) - g_i(x_j)| < \frac{\varepsilon}{3}$$

and so $n, m \geq N$ that

$$\begin{aligned} |g_n(x) - g_m(x)| &\leq |g_n(x) - g_n(x_j)| + |g_n(x_j) - g_m(x_j)| + |g_m(x_j) - g_m(x)| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon. \end{aligned}$$

□

Corollary 20.21. *Suppose X is a compact metric space. Let $S \subset \mathscr{C}(X)$ be a closed, bounded and equicontinuous set. Then S is compact.*

Corollary 20.22. *Suppose (f_n) is a sequence of differentiable functions on $[a,b]$, (f_n') is uniformly bounded, and there exists $x_0 \in [a,b]$ such that $(f_n(x_0))$ is bounded. Then there exists a uniformly convergent subsequence (f_{n_k}).*

20.5 Stone–Weierstrass Approximation Theorem

Perhaps surprisingly, even a very badly behaving continuous function is really just a uniform limit of polynomials. We cannot really get any "nicer" as a function than a polynomial.

Weierstrass's Version

Theorem 20.23 (Weierstrass approximation theorem). *If $f\colon [a,b] \to \mathbb{C}$ is continuous, there exists a sequence of polynomials (P_n) such that $P_n \rightrightarrows f$ on $[a,b]$.*
If f is real, then P_n may be taken real.

Proof. WLOG assume that $[a,b] = [0,1]$. We may also assume that $f(0) = f(1) = 0$. For if the theorem is proved for this case, consider

$$g(x) = f(x) - f(0) - x[f(1) - f(0)] \quad (0 \le x \le 1).$$

Here $g(0) = g(1) = 0$, and if g can be obtained as the limit of a uniformly convergent sequence of polynomials, it is clear that the same is true for f, since $f - g$ is a polynomial.

Furthermore, we define $f(x)$ to be zero for x outside $[0,1]$. Then f is uniformly continuous on the whole line.

Let

$$Q_n(x) = c_n(1-x^2)^n \quad (n = 1,2,\dots),$$

where c_n is chosen such that

$$\int_{-1}^{1} Q_n(x)\,\mathrm{d}x = 1 \quad (n = 1,2,\dots).$$

We need some information about the order of magnitude of c_n. Since

$$\begin{aligned}
\int_{-1}^{1} (1-x^2)^n\,\mathrm{d}x &= 2\int_{0}^{1} (1-x^2)^n\,\mathrm{d}x \\
&\ge 2\int_{0}^{\frac{1}{\sqrt{n}}} (1-x^2)^n\,\mathrm{d}x \\
&\ge 2\int_{0}^{\frac{1}{\sqrt{n}}} (1-nx^2)\,\mathrm{d}x \\
&= \frac{4}{3\sqrt{n}} \\
&> \frac{1}{\sqrt{n}},
\end{aligned}$$

it follows from (48) that

$$c_n < \sqrt{n}.$$

The inequality $(1-x^2)^n \geq 1-nx^2$ which we used above is easily shown to be true by considering the function

$$(1-x^2)^n - 1 + nx^2$$

which is zero at $x = 0$ and whose derivative is positive in $(0,1)$.

For any $\delta > 0$, (49) implies

$$Q_n(x) \leq \sqrt{n}(1-\delta^2)^n \quad (\delta \leq |x| \leq 1),$$

so that $Q_n \rightrightarrows 0$ in $\delta \leq |x| \leq 1$.

Now let

$$P_n(x) = \int_{-1}^{1} f(x+t)Q_n(t)\,\mathrm{d}t \quad (0 \leq x \leq 1).$$

Our assumptions about f show, by a simple change of variable, that

$$P_n(x) = \int_{-x}^{1-x} f(x+t)Q_n(t)\,\mathrm{d}t = \int_0^1 f(t)Q_n(t-x)\,\mathrm{d}t\,,$$

and the last integral is clearly a polynomial in x. Thus (P_n) is a sequence of polynomials, which are real if f is real.

Given $\varepsilon > 0$, we choose $\delta > 0$ such that

$$|y-x| < \delta \implies |f(y)-f(x)| < \frac{\varepsilon}{2}.$$

Let $M = \sup|f(x)|$, Using (48), (50), and the fact that $Q_n(x) \geq 0$, we see that for $0 \leq x \leq 1$,

$$\begin{aligned}
|P_n(x) - f(x)| &= \left| \int_{-1}^{1} [f(x+t)-f(x)]Q_n(t)\,\mathrm{d}t \right| \\
&\leq \int_{-1}^{1} |f(x+t)-f(x)|Q_n(t)\,\mathrm{d}t \\
&\leq 2M \int_{-1}^{-\delta} Q_n(t)\,\mathrm{d}t + \frac{\varepsilon}{2}\int_{-\delta}^{\delta} Q_n(t)\,\mathrm{d}t + 2M\int_{\delta}^{1} Q_n(t)\,\mathrm{d}t \\
&\leq 4M\sqrt{n}(1-\delta^2)^n + \frac{\varepsilon}{2} \\
&< \varepsilon
\end{aligned}$$

for all large enough n, which proves the theorem. □

Think about the consequences of the theorem. If you have any property that gets preserved under uniform convergence and it is true for polynomials, then it must be true for all continuous functions.

Let us note an immediate application of the Weierstrass theorem. We have already seen that countable dense subsets can be very useful.

Corollary 20.24. *The metric space $\mathscr{C}([a,b],\mathbb{C})$ contains a countable dense subset.*

Corollary 20.25. *For every interval $[-a,a]$, there exists a sequence of real polynomials P_n such that $P_n(0)=0$ and*

$$\lim_{n\to\infty} P_n(x) = |x|$$

uniformly on $[-a,a]$.

Algebra of Functions

We shall now isolate those properties of the polynomials which make the Weierstrass theorem possible.

Definition 20.26. A family $\mathscr{A}$ of complex-valued functions $f\colon X\to\mathbb{C}$ is an ***algebra*** if, for all $f,g\in\mathscr{A}$, $c\in\mathbb{C}$,

(i) $f+g\in\mathscr{A}$; (closed under addition)

(ii) $fg\in\mathscr{A}$; (closed under multiplication)

(iii) $cf\in\mathscr{A}$. (closed under scalar multiplication)

If we talk of an algebra of real-valued functions, then of course we only need the above to hold for $c\in\mathbb{R}$.
$\mathscr{A}$ is *uniformly closed* if the limit of every uniformly convergent sequence in $\mathscr{A}$ is also in $\mathscr{A}$.
Let $\mathscr{B}$ be the set of all limits of uniformly convergent sequences in $\mathscr{A}$. Then $\mathscr{B}$ is the *uniform closure* of $\mathscr{A}$.

Example.

- $\mathscr{C}(X,Y)$ is an algebra of functions.

Proposition 20.27. *Let $\mathscr{B}$ be the uniform closure of an algebra $\mathscr{A}$ of bounded functions. Then $\mathscr{B}$ is a uniformly closed algebra.*

Now let us distill the right properties of polynomials that were sufficient for an approximation theorem.

Definition 20.28. Let $\mathscr{A}$ be a family of functions defined on X.
We say $\mathscr{A}$ *separates points* if for every $x, y \in X$, with $x \neq y$ there exists $f \in \mathscr{A}$ such that $f(x) \neq f(y)$.
We say $\mathscr{A}$ *vanishes at no point* if for every $x \in X$ there exists $f \in \mathscr{A}$ such that $f(x) \neq 0$.

Example.

Proposition 20.29. *Suppose $\mathscr{A}$ is an algebra of functions on X, that separates points and vanishes at no point. Suppose x, y are distinct points of X and $c, d \in \mathbb{C}$. Then there exists $f \in \mathscr{A}$ such that*

$$f(x) = c, \quad f(y) = d.$$

The Theorem

We now have all the material needed for Stone's generalisation of the Weierstrass theorem.

Theorem 20.30 (Stone–Weierstrass approximation theorem)**.** *Let X be a compact metric space and $\mathscr{A}$ an algebra of real-valued continuous functions on X, such that $\mathscr{A}$ separates points and vanishes at no point. Then the uniform closure of $\mathscr{A}$ is all of $\mathscr{C}(X, \mathbb{R})$.*

Exercises

21 Some Special Functions

21.1 Power Series

Definition 21.1. Given a sequence (a_n) of complex numbers, a ***power series*** takes the form

$$\sum_{n=0}^{\infty} a_n z^n,$$

where $z \in \mathbb{C}$; the numbers a_n are called the *coefficients* of the series.

More generally, we may consider series

$$\sum_{n-0}^{\infty} a_n (z - z_0)^n$$

which are power series with respect to the centre z_0, but the difference is so slight that we need not do so in a formal manner.

Example. Consider the *geometric series*

$$1 + z + z^2 + \cdots + z^n + \cdots$$

whose partial sums can be written in the form

$$\frac{1 - z^n}{1 - z}.$$

Since $z^n \to 0$ for $|z| < 1$, and $|z^n| \geq 1$ for $|z| \geq 1$, we conclude that the geometric series converges to $1/(1-z)$ for $|z| < 1$, and diverges for $|z| \geq 1$.

It turns out that the behavior of the geometric series is typical. Indeed, we shall find that every power series converges inside a circle and diverges outside the same circle (except that it may happen that the series converges only for $z = 0$, or that it converges for all values of z).

Lemma 21.2 (Cauchy–Hadamard theorem). *Given the power series $\sum a_n z^n$, define*

$$\frac{1}{R} = \limsup_{n\to\infty} \sqrt[n]{|a_n|}. \tag{21.1}$$

(i) If $|z| < R$, $\sum a_n z^n$ converges.

(ii) If $|z| > R$, $\sum a_n z^n$ diverges.

We call R the ***radius of convergence*** of $\sum a_n z^n$, and call the circle $|z| = R$ the *circle* or ***disk of convergence***; nothing is claimed about the convergence on the circle.

Proof. Let $b_n = a_n z^n$. We apply the root test:

$$\limsup_{n\to\infty} \sqrt[n]{|b_n|} = \limsup_{n\to\infty} \sqrt[n]{|a_n z^n|} = |z| \limsup_{n\to\infty} \sqrt[n]{|a_n|} = \frac{|z|}{R}.$$

(i) If $|z| < R$, then $\limsup_{n\to\infty} \sqrt[n]{|b_n|} < 1$. By the root test, $\sum a_n z^n$ converges absolutely and thus converges.

(ii) If $|z| > R$, then $\limsup_{n\to\infty} \sqrt[n]{|b_n|} > 1$. By the root test, $\sum a_n z^n$ diverges.

□

Example.

- $\sum n^n z^n$ has $R = 0$.
- $\sum \frac{z^n}{n!}$ has $R = +\infty$. (In this case the ratio test is easier to apply than the root test.)
- The geometric series $\sum z^n$ has $R = 1$; if $|z| = 1$, the series diverges, since (z^n) does not tend to 0 as $n \to \infty$.
- $\sum \frac{z^n}{n}$ has $R = 1$; it diverges if $z = 1$, and converges for all other z with $|z| = 1$.
- $\sum \frac{z^n}{n^2}$ has $R = 1$; it converges for all z with $|z| = 1$, by the comparison test, since $\left|\frac{z^n}{n^2}\right| = \frac{1}{n^2}$.

In the previous result, we have shown that the radius of convergence can be found by using the root test. We can also find it using the ratio test (which is easier to compute).

Lemma 21.3. *If $\sum a_n z^n$ has radius of convergence R, then*

$$R = \lim_{n\to\infty} \left|\frac{a_n}{c_{n+1}}\right|,$$

if this limit exists.

Proof. By the ratio test, $\sum a_n z^n$ converges if

$$\lim_{n\to\infty}\left|\frac{a_{n+1}z^{n+1}}{a_n z^n}\right| < 1.$$

This is equivalent to

$$|z| < \frac{1}{\lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right|} = \lim_{n\to\infty}\left|\frac{a_n}{a_{n+1}}\right|.$$

□

Proposition 21.4. *Suppose the radius of convergence of $\sum c_n z^n$ is 1, and suppose $c_0 \geq c_1 \geq c_2 \geq \cdots$, $c_n \to 0$. Then $\sum c_n z^n$ converges at every point on the circle $|z| = 1$, except possibly at $z = 1$.*

Proof. Let

$$a_n = z^n, \quad b_n = c_n.$$

Then the hypothesis of 16.37 are satisfied, since

$$|A_n| = \left|\sum_{k=0}^{n} z^k\right| = \left|\frac{1-z^{n+1}}{1-z}\right| \leq \frac{2}{|1-z|}$$

if $|z| = 1$, $z \neq 1$. □

Definition 21.5. An ***analytic function*** is a function that can be represented by a power series; that is, functions of the form

$$f(z) = \sum_{n=0}^{\infty} a_n z^n.$$

We shall restrict ourselves to real values of z (since we have yet to define complex differentiation). Instead of circles of convergence, we shall therefore encounter intervals of convergence.

If $\sum a_n x^n$ converges for all $x \in (-R, R)$, for some $R > 0$, we say that f is *expanded in a power series* about the point $x = 0$.

Proposition 21.6. *Suppose $\sum a_n x^n$ converges for $|x| < R$. Let*

$$f(x) = \sum_{n=0}^{\infty} a_n x^n \quad (|x| < R).$$

Then

(i) $\sum a_n x^n$ converges absolutely and uniformly on $(-r, r)$ where $r < R$;

(ii) $f(x)$ is continuous and differentiable on $(-R,R)$, and

$$f'(x) = \sum_{n=1}^{\infty} n a_n x^{n-1} \quad (|x| < R).$$

Proof.

(i) We will show that $\sum a_n x^n$ converges absolutely and uniformly on $[-R+\varepsilon, R-\varepsilon]$ for all $\varepsilon > 0$.

Idea. Weierstrass M-test.

Let $\varepsilon > 0$ be given. For $|x| \le R - \varepsilon$, notice that we have

$$|a_n x^n| = |a_n||x|^n \le |a_n|(R-\varepsilon)^n$$

for each $n = 1, 2, \ldots$. Consider the series

$$\sum |a_n|(R-\varepsilon)^n.$$

Since every power series converges (absolutely) in the interior of its internal of convergence (by 21.2), $\sum |a_n|(R-\varepsilon)^n$ converges.

By the Weierstrass M-test, $\sum a_n x^n$ uniformly converges on $[-R+\varepsilon, R-\varepsilon]$.

(ii) Since $\lim_{n\to\infty} \sqrt[n]{n} = 1$, we have

$$\limsup_{n\to\infty} \sqrt[n]{n|a_n|} = \limsup_{n\to\infty} \sqrt[n]{|a_n|},$$

so the series $\sum_{n=0}^{\infty} a_n x^n$ and $\sum_{n=1}^{\infty} n a_n x^{n-1}$ have the same radius of convergence; thus $\sum_{n=1}^{\infty} n a_n x^{n-1}$ has radius of convergence R.

Idea. Interchange limits and derivatives.

Since $\sum_{n=1}^{\infty} n a_n x^{n-1}$ is a power series, by (i), it converges uniformly in $[-R+\varepsilon, R-\varepsilon]$, for every $\varepsilon > 0$.

Consider the partial sums $f_n(x) = \sum_{k=0}^{n} a_k x^k$; evidently f_n are differentiable on $[-R+\varepsilon, R-\varepsilon]$, and $(f_n(x))$ converges whenever $|x| < R$. Also,

$$f_n'(x) = \sum_{k=1}^{n} k a_k x^{k-1},$$

converge uniformly on $[-R+\varepsilon, R-\varepsilon]$. By 20.14, for all $|x| < R$,

$$f'(x) = \lim_{n\to\infty} f_n'(x) = \lim_{n\to\infty} \sum_{k=1}^{n} k a_k x^{k-1} = \sum_{n=1}^{\infty} n a_n x^{n-1}.$$

Since f is differentiable on $(-R,R)$, by 18.2, f is continuous on $(-R,R)$.

□

Corollary 21.7. *f is infinitely differentiable in $(-R,R)$; its derivatives are given by*

$$f^{(k)}(x) = \sum_{n=k}^{\infty} n(n-1)\cdots(n-k+1) a_n x^{n-k}. \tag{21.2}$$

In particular,

$$f^{(k)}(0) = k! a_k \quad (k = 0,1,2,\dots). \tag{21.3}$$

Proof. Apply the previous result successively to $f, f', f'', \dots$.

Then plug in $x = 0$. □

Remark. (21.3) is very interesting.

- It shows, on the one hand, that the coefficients of the power series development of f are determined by the values of f and its derivatives at a single point.
- On the other hand, if the coefficients are given, the values of the derivatives of f at the center of the interval of convergence can be read off immediately from the power series.

If the series converges at an endpoint, say at $x = R$, then f is continuous not only in $(-R,R)$, but also at $x = R$, as shown by the following result (for simplicity of notation, we take $R = 1$).

Theorem 21.8 (Abel's theorem). *Suppose $\sum a_n$ converges. Let*

$$f(x) = \sum_{n=0}^{\infty} a_n x^n \quad (-1 < x < 1).$$

Then $f(x)$ is continuous at $x = 1$.

Proof. We want to show that $\lim_{x\to 1} f(x) = \sum_{n=0}^{\infty} a_n$.

Let

$$s_n = a_0 + \cdots + a_n, \quad s_{-1} = 0.$$

Then we write

$$\begin{aligned}\sum_{n=0}^{m} a_n x^n &= \sum_{n=0}^{m} (s_n - s_{n-1})x^n \\ &= \sum_{n=0}^{m} s_n x^n - \sum_{n=1}^{m} s_{n-1}x^n \\ &= \sum_{n=0}^{m} s_n x^n - \sum_{n=0}^{m-1} s_n x^{n+1} \\ &= (1-x)\sum_{n=0}^{m-1} s_n x^n + s_m x^m.\end{aligned}$$

For $|x| < 1$, we let $m \to \infty$ and obtain

$$f(x) = (1-x)\sum_{n=0}^{\infty} s_n x^n.$$

Suppose $s_n \to s$. We will show that $\lim_{x\to 1} f(x) = s$. Fix $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that

$$n \geq N \implies |s_n - s| < \frac{\varepsilon}{2}.$$

Note that for $|x| < 1$, since $\sum_{n=0}^{\infty} x^n = \dfrac{1}{1-x}$, we can write

$$(1-x)\sum_{n=0}^{\infty} s x^n = s.$$

If $x > 1-\delta$, for some suitably chosen $\delta > 0$, we have

$$\begin{aligned}|f(x) - s| &= \left|(1-x)\sum_{n=0}^{\infty}(s_n - s)x^n\right| \\ &= (1-x)\left|\sum_{n=0}^{N}(s_n - s)x^n + \sum_{n=N+1}^{\infty}(s_n - s)x^n\right| \\ &\leq (1-x)\left|\sum_{n=0}^{N}(s_n - s)x^n\right| + (1-x)\sum_{n=N+1}^{\infty}(s_n - s)x^n.\end{aligned}$$

Note that

$$\begin{aligned}(1-x)\left|\sum_{n=N+1}^{\infty}(s_n-s)x^n\right| &\le (1-x)\sum_{n=N+1}^{\infty}|s_n-s|\,|x|^n\\ &< \frac{\varepsilon}{2}(1-x)\sum_{n=N+1}^{\infty}x^n\\ &= \frac{\varepsilon}{2}(1-x)\frac{x^{N+1}}{1-x} < \frac{\varepsilon}{2}\end{aligned}$$

and

$$\begin{aligned}(1-x)\left|\sum_{n=0}^{N}(s_n-s)x^n\right| &\le (1-x)\sum_{n=0}^{N}|s_n-s|\,|x|^n\\ &< (1-x)\sum_{n=0}^{N}|s_n-s|\end{aligned}$$

which can be bounded by, say, M because there are only finitely many terms in the sum. Choosing $\delta < \frac{\varepsilon}{2M}$ gives

$$(1-x)\sum_{n=0}^{N}|s_n-s| < (1-(1-\delta))\sum_{n=0}^{N}|s_n-s| < \delta M < \frac{\varepsilon}{2}.$$

Hence

$$|f(x)-s| < \frac{\varepsilon}{2}+\frac{\varepsilon}{2} = \varepsilon,$$

and therefore $\lim_{x\to 1} f(x) = s$, as desired. □

We now require a result concerning an inversion in the order of summation.

Proposition 21.9 (Fubini's theorem for sums). *Given a double sequence* (a_{ij}), $i = 1,2,\ldots$, $j = 1,2,\ldots$, *suppose that*

$$\sum_{j=1}^{\infty}|a_{ij}| = b_i \quad (i = 1,2,\ldots)$$

and $\sum b_i$ *converges. Then*

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}a_{ij} = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty}a_{ij}. \tag{21.4}$$

Remark. This is analogous to Fubini's theorem for the swapping of double integrals.

Proof. Let E be a countable set:

$$E = \{x_0, x_1, x_2, \ldots\},$$

and suppose $x_n \to x_0$ as $n \to \infty$. Define the sequence of functions $f_i : E \to \mathbb{C}$ by

$$f_i(x_0) = \sum_{j=1}^{\infty} a_{ij} \quad (i = 1,2,\dots)$$
$$f_i(x_n) = \sum_{j=1}^{n} a_{ij} \quad (i,n = 1,2,\dots)$$

See that each f_i is continuous at x_0.

Since $|f_i(x)| \leq b_i$ for $x \in E$ (by triangle inequality), and $\sum b_i$ converges, by the Weierstrass M-test, $\sum_{i=1}^{n} f_i(x)$ converges uniformly. Let

$$g(x) = \sum_{i=1}^{\infty} f_i(x) \quad (x \in E).$$

By 7.11, g is continuous at x_0, so

$$\lim_{n\to\infty} g(x_n) = g(x_0).$$

It follows that

$$\begin{aligned}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{ij} &= \sum_{i=1}^{\infty} f_i(x_0) = g(x_0) = \lim_{n\to\infty} g(x_n) \\ &= \lim_{n\to\infty} \sum_{i=1}^{\infty} f_i(x_n) = \lim_{n\to\infty} \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{ij} \\ &= \lim_{n\to\infty} \sum_{j=1}^{n}\sum_{i=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} a_{ij}.\end{aligned}$$

□

Theorem 21.10 (Taylor's theorem). *Suppose $\sum c_n x^n$ converges in $|x| < R$, let*

$$f(x) = \sum_{n=0}^{\infty} c_n x^n.$$

If $a \in (-R,R)$, then f can be expanded in a power series about the point $x = a$ which converges in $|x-a| < R - |a|$, and

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n \quad (|x-a| < R - |a|). \tag{21.5}$$

Proof. We have

$$\begin{aligned} f(x) &= \sum_{n=0}^{\infty} c_n x^n = \sum_{n=0}^{\infty} c_n \left[(x-a)+a\right]^n \\ &= \sum_{n=0}^{\infty} c_n \sum_{m=0}^{n} \binom{n}{m} a^{n-m}(x-a)^m \\ &= \sum_{m=0}^{\infty} \left[\sum_{n=m}^{\infty} \binom{n}{m} c_n a^{n-m}\right] (x-a)^m \end{aligned} \tag{1}$$

This is the desired expansion about the point $x = a$. We need to show that the swapping of summations in (1) is valid, which is applicable only if $\binom{n}{m} c_n a^{n-m}(x-a)^m$ satisfies Theorem 8.3, i.e.

$$\sum_{n=0}^{\infty} \sum_{m=0}^{n} \left| c_n \binom{n}{m} a^{n-m}(x-a)^m \right|$$

converges. Write

$$\begin{aligned} &\sum_{n=0}^{\infty} \sum_{m=0}^{n} \left| c_n \binom{n}{m} a^{n-m}(x-a)^m \right| \\ &= \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \binom{n}{m} |c_n||a|^{n-m}|x-a|^m \\ &= \sum_{n=0}^{\infty} |c_n| \left(|x-a|+|a|\right)^n, \end{aligned}$$

which converges if and only if $|x-a|+|a| < R$.

Finally, the form of the coefficients in (21.5) follows from (21.3): differentiate $f(x)$ repeatedly to obtain

$$\begin{aligned} f^{(m)}(x) &= \sum_{n=m}^{\infty} c_n n(n-1)\cdots(n-m-1)x^{n-m} \\ &= \sum_{n=m}^{\infty} c_n m! \binom{n}{m} x^{n-m} \end{aligned}$$

and then plug in $x = a$. □

If two power series converge to the same function in $(-R,R)$, (7) shows that the two series must be identical, i.e., they must have the same coefficients. It is interesting that the same conclusion can be deduced from much weaker hypotheses:

Proposition 21.11. *Suppose $\sum a_n x^n$ and $\sum b_n x^n$ converge in $S = (-R,R)$. Let E be the set of all $x \in S$ such that*

$$\sum_{n=0}^{\infty} a_n x^n = \sum_{n=0}^{\infty} b_n x^n.$$

If E has a limit point in S, then $a_n = b_n$ for $n = 0, 1, 2, \ldots$. Hence (20) holds for all $x \in S$.

Proof. Let $c_n = a_n - b_n$, and let

$$f(x) = \sum_{n=0}^{\infty} c_n x^n \quad (x \in S)$$

We will show that $c_n = 0$, so that $f(x) = 0$ on E.

Let A be the set of all limit points of E in S, and let B consist of all other points of S. It is clear from the definition of "limit point" that B is open. Suppose we can prove that A is open. Then A and B are disjoint open sets. Hence they are separated (Definition 2.45). Since $S = A \cup B$, and S is connected, one of A and B must be empty. By hypothesis, A is not empty. Hence B is empty, and $A = S$. Since f is continuous in S, $A \subset E$. Thus $E = S$, and (7) shows that $c_n = 0$ for n = 0, 1, 2, ... , which is the desired conclusion.

Thus we have to prove that A is open. If $x_0 \in A$, Theorem 8.4 shows that

$$f(x) = \sum_{n=0}^{\infty} d_n (x - x_0)^n \quad (|x - x_0| < R - |x_0|).$$

We claim that $d_n = 0$ for all n. Otherwise, let k be the smallest nonnegative integer such that $d_k \neq 0$. Then

$$f(x) = (x - x_0)^k g(x) \quad (|x - x_0| < R - |x_0|),$$

where

$$g(x) = \sum_{m=0}^{\infty} d_{k+m} (x - x_0)^m.$$

Since g is continuous at x_0 and

$$g(x_0) = d_k \neq 0,$$

there exists $\delta > 0$ such that $g(x) \neq 0$ if $|x - x_0| < \delta$. It follows from (23) that $f(x) \neq 0$ if $0 < |x - x_0| < \delta$. But this contradicts the fact that x_0 is a limit point of E.

Thus $d_n = 0$ for all n, so that $f(x) = 0$ for all x for which (22) holds, i.e., in a neighborhood of x_0. This shows that A is open, and completes the proof. to do

□

Exponential Function

Definition 21.12 (Exponential function)**.** For $z \in \mathbb{C}$, define

$$\exp(z) := \sum_{n=0}^{\infty} \frac{z^n}{n!}. \tag{21.6}$$

Notation. We shall usually replace $\exp(z)$ by the customary shorter expression e^z.

The next result tells us that $\exp(z)$ does have a value for each $z \in \mathbb{C}$ (since the series does not diverge).

Lemma 21.13. $\exp(z)$ *converges for every* $z \in \mathbb{C}$.

Proof. We have

$$\lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right| = \lim_{n\to\infty} \left| \frac{z}{n+1} \right| = |z| \lim_{n\to\infty} \frac{1}{n+1} = 0 < 1.$$

By the ratio test, the series converges absolutely for all $z \in \mathbb{C}$, and thus converges for all $z \in \mathbb{C}$. □

Hence $\exp(z)$ has infinite radius of convergence.

The series converges uniformly on every bounded subset of the complex plane. Thus exp is a continuous function.

Lemma 21.14 (Addition formula)**.** *For* $z, w \in \mathbb{C}$,

$$e^{z+w} = e^z e^w. \tag{21.7}$$

Proof. By multiplication of absolutely convergent series,

$$\begin{aligned} e^z e^w &= \sum_{n=0}^{\infty} \frac{z^n}{n!} \sum_{m=0}^{\infty} \frac{w^m}{m!} \\ &= \sum_{k=0}^{\infty} \left(\frac{z^k}{k!} + \frac{z^{k-1}}{(k-1)!}\frac{w}{1!} + \cdots + \frac{w^k}{k!} \right) \\ &= \sum_{k=0}^{\infty} \frac{1}{k!} \sum_{m+n=k} \binom{k}{n} z^n w^{k-n} \\ &= \sum_{k=0}^{\infty} \frac{1}{k!} (z+w)^k \\ &= e^{z+w} \end{aligned}$$

□

A special case of the addition formula

$$e^z e^{-z} = 1 \tag{21.8}$$

shows that $e^z \neq 0$ for all $z \in \mathbb{C}$.

Lemma 21.15. *The restriction of* exp *to* $\mathbb{R}$ *is a monotonically increasing positive function, and*

$$\lim_{x\to\infty} e^x = \infty, \quad \lim_{x\to-\infty} e^x = 0.$$

Proof. By (21.6), $\exp(x) > 0$ if $x > 0$ (since all the terms are positive); hence (21.8) shows that $\exp(x) > 0$ for all real x.

By (21.6), $\exp(x) \to +\infty$ as $x \to +\infty$; hence (21.8) shows that $\exp(x) \to 0$ as $x \to -\infty$ along the real axis.

By (21.6), $0 < x < y$ implies that $\exp(x) < \exp(y)$; by (21.8), it follows that $\exp(-y) < \exp(-x)$. □

Lemma 21.16. $(e^x)' = e^x$ *for all* $x \in \mathbb{R}$.

Proof. By the addition formula,

$$(e^x)' = \lim_{h\to0} \frac{e^{x+h} - e^x}{h} = \lim_{h\to0} \frac{e^x e^h - e^x}{h} = e^x \lim_{h\to0} \frac{e^h - 1}{h} = e^x.$$

□

Since exp is strictly increasing and differentiable on $\mathbb{R}$, it has an inverse function log which is also strictly increasing and differentiable and whose domain is $\exp(\mathbb{R}) = \mathbb{R}^+$. Define the ***logarithm function*** log by

$$\exp(\log(y)) = y \quad (y > 0),$$

or, equivalently, by

$$\log(\exp(x)) = x \quad (x \in \mathbb{R}).$$

Differentiating this using the chain rule, we obtain

$$\log'(\exp(x)) \exp(x) = 1.$$

Writing $y = \exp(x)$, this gives us

$$\log'(y) = \frac{1}{y} \quad (y > 0).$$

Taking $x = 0$, we see that $\log(1) = 0$. Hence

$$\log y = \int_1^y \frac{1}{x}\,\mathrm{d}x.$$

Lemma 21.17 (Addition formula). *If $u, v > 0$ then*

$$\log(uv) = \log(u) + \log(v). \tag{21.9}$$

Proof. Write $u = \exp(x)$, $v = \exp(y)$. Then

$$\begin{aligned}\log(uv) &= \log\left(\exp(x)\exp(y)\right)\\ &= \log\left(\exp(x+y)\right)\\ &= x+y\\ &= \log(u)+\log(v).\end{aligned}$$

□

Trigonometric Functions

Definition 21.18. For $z \in \mathbb{C}$, define

$$\cos z := \frac{e^{iz}+e^{-iz}}{2}, \quad \sin z := \frac{e^{iz}-e^{-iz}}{2i}. \tag{21.10}$$

Substitution into (21.6) shows that the trigonometric functions have the power series

$$\begin{aligned}\cos z &= \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!}\\ \sin z &= \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{(2n+1)!}\end{aligned}$$

For real z, they reduce to the familiar Taylor developments of $\cos x$ and $\sin x$, with the significant difference that we have now redefined these functions without use of geometry.

From (21.10) we obtain *Euler's formula*

$$e^{iz} = \cos z + i \sin z.$$

as well as the Pythagoras theorem

$$\cos^2 z + \sin^2 z = 1.$$

Euler's formula allows us to calculate the value of the exponential function for all $z \in \mathbb{C}$:

$$e^z = e^{x+iy} = e^x e^{iy} = e^x(\cos y + i \sin y).$$

Let us now restrict ourselves to considering real values $x \in \mathbb{R}$. Then $\cos x$ and $\sin x$ are real for real x. By Euler's identity, $\cos x$ and $\sin x$ are the real and imaginary parts, respectively, of e^{ix}.

Lemma 21.19 (Derivative of trigonometric functions). *For $x \in \mathbb{R}$,*

$$\cos' x = -\sin x, \quad \sin' x = \cos x.$$

The addition formulae

$$\cos(a+b) = \cos a \cos b - \sin a \sin b$$
$$\sin(a+b) = \cos a \sin b + \sin a \cos b$$

are direct consequences of (21.10) and the addition formula for the exponential function.

The other trigonometric functions $\tan z$, $\cot z$, $\sec z$, $\csc z$ are of secondary importance. They are defined in terms of $\cos z$ and $\sin z$ in the customary manner. We find for instance

$$\tan z = -i \frac{e^{iz} - e^{-iz}}{e^{iz} + e^{-iz}}.$$

Observe that all the trigonometric functions are rational functions of e^{iz}.

Let x_0 be the smallest positive number such that $\cos x_0 = 0$. This exists, since the set of zeros of a continuous function is closed, and $\cos 0 \neq 0$. We define the number

$$\pi := 2x_0.$$

Then $\cos \frac{\pi}{2} = 0$, and $\sin \frac{\pi}{2} = \pm 1$. Since

$$\sin' x_0 = \cos x_0 > 0$$

in $\left(0, \frac{\pi}{2}\right)$ and since $\sin 0 = 0$, we have $\sin x_0 > 0$; hence $\sin \frac{\pi}{2} = 1$.

Lemma 21.20. exp *is periodic, with period $2\pi i$.*

Proof. By Euler's formula,

$$e^{\pi i/2} = \cos \frac{\pi}{2} + i \sin \frac{\pi}{2} = i.$$

By the addition formula, $e^{\pi i} = i^2 = -1$, $e^{2\pi i} = (-1)^2 = 1$. Hence for all $z \in \mathbb{C}$,

$$e^{z+2\pi i} = e^z e^{2\pi i} = e^z.$$

□

Since cos and sin are defined in terms of e^{iz}, it follows that cos and sin are periodic, with period 2π.

Lemma 21.21. *The mapping $t \mapsto e^{it}$ maps the real axis onto the unit circle.*

Proof. We already know that $t \mapsto e^{it}$ maps the real axis *into* the unit circle.

Fix $w \in \mathbb{C}$, where $|w| = 1$. We will show that $w = e^{it}$ for some real t. Write $w = u + iv$ for some real u, v. We shall separately consider the cases where w lies in each of the four quadrants of the Argand diagram. (Case n corresponds to the n-th quadrant, for $1 \leq n \leq 4$.)

Case 1: $u \geq 0$ and $v \geq 0$. Since $u \leq 1$, the definition of π shows that there exists $t \in [0, \pi/2]$ such that $\cos t = u$.

By Pythagoras' theorem,

$$\sin^2 t = 1 - \cos^2 t = 1 - u^2 = v^2$$

where the last equality follows from $|w| = u^2 + v^2 = 1$. Since $\sin t \geq 0$ for all $t \in [0, \pi/2]$, we have $\sin t = v$.

Hence we have $w = \cos t + i \sin t$. By Euler's formula, $w = e^{it}$.

Case 2: $u < 0$ and $v \geq 0$. The preceding conditions are satisfied by $-iw$. Hence $-iw = e^{it}$ for some real t, and $w = e^{i(t+\pi/2)}$.

(Intuitively, rotate w clockwise such that it lies in the first quadrant.)

Cases 3 and 4: $v < 0$. The preceding two cases show that $-w = e^{it}$ for some real t. Hence $w = e^{i(t+\pi)}$.

(Intuitively, reflect w over the real axis such that it lies in the first or second quadrant.)

This completes the proof. □

In particular, for every $w \in \mathbb{C}$ with $|w| = 1$, there is one and only one $t \in [0, 2\pi)$ such that $w = e^{it}$.

Remark. From an algebraic point of view, the mapping $w = e^{it}$ establishes a homomorphism from $(\mathbb{R}, +)$ to $\mathbb{T}$ (where $\mathbb{T}$ denotes the multiplicative group of complex numbers with absolute value 1).

The kernel of the homomorphism is the subgroup formed by all integral multiples $2\pi n$.

Lemma 21.22. *If $w \in \mathbb{C}$ and $w \neq 0$, then $w = e^z$ for some $z \in \mathbb{C}$.*

Proof. Since $w \neq 0$, we can define $\alpha = w/|w|$. Then $w = |w|\alpha$.

By 21.15, there is a real x such that $|w| = e^x$.

Since $|\alpha| = 1$, the previous result shows that $\alpha = e^{iy}$ for some real y.

Hence $w = e^{x+iy}$, so take $z = x + iy$ and we are done. □

Hyperbolic Functions

The hyperbolic cosine and sine functions are defined by

$$\cosh z := \frac{e^z + e^{-z}}{2}, \quad \sinh z := \frac{e^z - e^{-z}}{2}.$$

The other hyperbolic functions are given by

$$\tanh z = \frac{\sinh z}{\cosh z}, \quad \coth z = \frac{\cosh z}{\sinh z}.$$

Some properties: One has

$$\cosh^2 z - \sinh^2 z = 1, \quad (\cosh z)' = \sinh z, \quad (\sinh z)' = \cosh z.$$

Logarithmic Function

We define the ***logarithm function*** log as the inverse of the exponential function. That is,

$$e^{\log w} = \log e^w = w.$$

Notation. Some people prefer to denote the logarithm function by ln instead.

By definition, $z = \log w$ is a root of the equation $e^z = w$. Since $e^z \neq 0$ for all $z \in \mathbb{C}$, $\log 0$ is not defined. For $w \neq 0$, let $z = x + iy$, then $e^{x+iy} = w$ is equivalent to

$$\begin{aligned} &(1) \quad e^x = |w| \\ &(2) \quad e^{iy} = \frac{w}{|w|}. \end{aligned}$$

(1) has a unique solution $x = \log |w|$, the *real logarithm* of the positive number $|w|$.

The RHS of (2) is a complex number of absolute value 1, so (2) has one and only one solution $y \in [0, 2\pi)$; (2) is also satisfied by all y that differ from this solution by an integral multiple of 2π.

We call the imaginary part of $\log w$ the *argument* of w, denoted as $\arg w$. (It is interpreted geometrically as the angle, measured in radians, from the positive real axis to the half line from 0 through the point w.) Then the argument is not unique – it has infinitely many values which

differ by multiples of 2π, and

$$\log w = \log|w| + i\arg w. \tag{21.11}$$

Again, note that $\log z$ is unique; it is defined up to multiples of $2\pi i$.

With a change of notation, if $|z| = r$ and $\arg z = \theta$, then

$$z = re^{i\theta}.$$

This notation is so convenient that it is used constantly, even when the exponential function is not otherwise involved.

Lemma 21.23.

$$\log z_1 z_2 = \log z_1 + \log z_2$$
$$\arg z_1 z_2 = \arg z_1 + \arg z_2$$

in the sense that both sides represent the same infinite set of complex numbers. If we want to compare a value on the left with a value on the right, then we can merely assert that they differ by a multiple of $2\pi i$ (or 2π).

Proof. This follows from the addition formula of the exponential function. □

Remark. The addition formula tells us that exp is a homomorphism from $\mathbb{C}$ to $\mathbb{C}^\times$. In particular, exp is surjective: if $w \in \mathbb{C}$, we can write

$$w = r(\cos\theta + i\sin\theta) = e^{\log r + i\theta}.$$

The kernel of exp consists of $z \in \mathbb{C}$ such that $e^z = 1$, or

$$e^{x+iy} = e^x(\cos y + i\sin y) = 1.$$

This implies that $x = 0$, and y is a multiple of 2π. Hence $\ker\exp = 2\pi i\mathbb{Z}$.

Also note that the restriction $\exp\colon \mathbb{R} \to \mathbb{R}^\times$ is an injective homomorphism, but it is not surjective.

Finally we discuss the inverse cosine which is obtained by solving the equation

$$\cos z = w.$$

By definition,

$$\frac{1}{2}(e^{iz} + e^{-iz}) = w.$$

This is a quadratic equation in e^{iz} with the roots

$$e^{iz} = w \pm \sqrt{w^2 - 1},$$

and consequently

$$z = \arccos w = -i \log\left(w \pm \sqrt{w^2 - 1}\right).$$

We can also write these values in the form

$$\arccos w = \pm i \log\left(w \pm \sqrt{w^2 - 1}\right)$$

for $w + \sqrt{w^2 - 1}$ and $w - \sqrt{w^2 - 1}$ are reciprocal numbers. The infinitely many values of $\arccos w$ reflect the evenness and periodicity of $\cos z$. The inverse sine is most easily defined by

$$\arcsin w = \frac{\pi}{2} - \arccos w.$$

It is worth emphasising that in the theory of complex analytic functions all elementary transcendental functions can thus be expressed through e^z and its inverse $\log z$. In other words, there is essentially only one elementary transcendental function.

21.2 Algebraic Completeness of the Complex Field

Let us prove the fundamental theorem of algebra, which states that the complex field is *algebraically complete* (every non-constant polynomial with complex coefficients has a complex root).

Theorem 21.24 (Fundamental Theorem of Algebra). *For $a_i \in \mathbb{C}$, let*

$$P(z) = \sum_{k=0}^{n} a_k z^k$$

where $n \geq 1$, $a_n \neq 0$. Then $P(z) = 0$ for some $z \in \mathbb{C}$.

Proof. WLOG assume $a_n = 1$ (then P is a monic polynomial). Let

$$\mu = \inf_{z \in \mathbb{C}} |P(z)|.$$

If $|z| = R$, then apply the triangle inequality to

$$z^n = P(z) - a_0 - a_1 z - \cdots - a_{n-1} z^{n-1}$$

to obtain

$$\begin{aligned} |z^n| &\leq |P(z)| + |a_0| + |a_1||z| + \cdots + |a_{n-1}||z|^{n-1} \\ R^n &\leq |P(z)| + |a_0| + |a_1|R + \cdots + |a_{n-1}|R^{n-1} \\ |P(z)| &\geq R^n \left(1 - |a_{n-1}|R^{-1} - \cdots - |a_0|R^{-n}\right). \end{aligned}$$

The RHS tends to ∞ as $R \to \infty$. Hence there exists R_0 such that $|P(z)| > \mu$ if $|z| > R_0$. Since $|P|$ is continuous on the closed disk $\overline{D}_{R_0}(0)$, Theorem 4.16 shows that $|P(z_0)| = \mu$ for some z_0.

Claim. $\mu = 0$.

Suppose otherwise, that $\mu \neq 0$.

Let $Q(z) = \dfrac{P(z+z_0)}{P(z_0)}$. Then Q is a non-constant polynomial, $Q(0) = 1$, and $|Q(z)| \geq 1$ for all z. There is a smallest integer k, $1 \leq k \leq n$ such that

$$Q(z) = 1 + b_k z^k + \cdots + b_n z^n \quad (b_k \neq 0).$$

By Theorem 8.7(d) there is a real θ such that

$$e^{ik\theta} b_k = -|b_k|.$$

If $r > 0$ and $r^k|b_k| < 1$, the above equation implies

$$\left|1 + b_k r^k e^{ik\theta}\right| = 1 - r^k|b_k|,$$

so that

$$\begin{aligned}\left|Q\left(re^{i\theta}\right)\right| &= |1 + b_k r^k e^{ik\theta} + b_{k+1} r^{k+1} e^{i(k+1)\theta} + \cdots + b_n r^n e^{in\theta}| \\ &\leq |1 + b_k r^k e^{ik\theta}| + |b_{k+1} r^{k+1} e^{i(k+1)\theta}| + \cdots + |b_n r^n e^{in\theta}| \\ &= 1 - r^k|b_k| + |b_{k+1}|r^{k+1} + \cdots + |b_n|r^n \\ &= 1 - r^k\left(|b_k| - r|b_{k+1}| - \cdots - r^{n-k}|b_n|\right).\end{aligned}$$

For sufficiently small r, the expression in braces is positive; hence $|Q(re^{i\theta})| < 1$, a contradiction. Thus $\mu = 0$, that is, $P(z_0) = 0$. □

21.3 Fourier Series

Definition 21.25. A ***trigonometric polynomial*** is a finite sum of the form

$$f(x) = a_0 + \sum_{n=1}^{N} (a_n \cos nx + b_n \sin nx) \quad (x \in \mathbb{R})$$

where $a_0, a_1 \ldots, a_N, b_1, \ldots, b_N \in \mathbb{C}$.

Using (21.10), we can write the above in the form

$$f(x) = \sum_{n=-N}^{N} c_n e^{inx}$$

for some constants $c_n \in \mathbb{C}$. This is a more convenient form of trigonometric polynomials, which we shall work with.

It is clear that every trigonometric polynomial is periodic, with period 2π.

For non-zero integer n, e^{inx} is the derivative of $\frac{1}{in} e^{inx}$, which also has period 2π. Hence

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{inx} \, \mathrm{d}x = \begin{cases} 1 & (n = 0) \\ 0 & (n = \pm 1, \pm 2, \ldots) \end{cases}$$

Definition 21.26. Let $f \in \mathscr{R}[-\pi, \pi]$. The ***Fourier coefficients*** of f are the numbers c_n, defined by

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-inx} \, \mathrm{d}x.$$

The series

$$\sum_{n=-\infty}^{\infty} c_n e^{inx}$$

formed with the Fourier coefficients is called the ***Fourier series*** of f; in this case we write

$$f \sim \sum_{n=-\infty}^{\infty} c_n e^{inx}.$$

We say f is an L^2 function if $|f|^2$ is Lebesgue integrable. The space of L^2 functions on a set E is denoted by $L^2(E)$. For all $f, g \in L^2(E)$, define the inner product

$$\langle f, g \rangle = \int_E f(x) \overline{g(x)} \, \mathrm{d}x.$$

Then the norm of f squared is defined as

$$\|f\|^2 := \langle f, f\rangle = \int_E |f(x)|^2 \,\mathrm{d}x.$$

We say that f and g are *orthogonal* if $\langle f, g\rangle = 0$.

Definition 21.27. Let (ϕ_n) be a sequence of complex functions on $[a,b]$.

(i) We say (ϕ_n) is an ***orthogonal system*** of functions on $[a,b]$ if $\langle \phi_n, \phi_m\rangle = 0$ for all $n \neq m$.

(ii) We say (ϕ_n) is an ***orthonormal system*** of functions on $[a,b]$ if (ϕ_n) is an orthogonal system, and $\|\phi_n\| = 1$ for all n.

Example.

- $\left\{\frac{1}{\sqrt{2\pi}} e^{inx}\right\}$ is an orthonormal system on $[-\pi, \pi]$.
- $\left\{\frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}}\cos nx, \frac{1}{\sqrt{\pi}}\sin nx\right\}$ is an orthonormal system on $[-\pi, \pi]$.

Proof. We have

$$\begin{aligned}
&\int_{-\pi}^{\pi} \cos nx \sin mx \,\mathrm{d}x \\
&= \int_{-\pi}^{\pi} \frac{e^{inx} + e^{-inx}}{2} \frac{e^{imx} - e^{-imx}}{2i} \,\mathrm{d}x \\
&= \int_{-\pi}^{\pi} \frac{e^{i(n+m)x} - e^{i(n-m)x} + e^{-i(n-m)x} - e^{-i(n+m)x}}{4i} \,\mathrm{d}x = 0
\end{aligned}$$

and similarly

$$\begin{aligned}
&\int_{-\pi}^{\pi} \cos nx \cos mx \,\mathrm{d}x \\
&= \int_{-\pi}^{\pi} \frac{e^{inx} + e^{-inx}}{2} \frac{e^{imx} + e^{-imx}}{2i} \,\mathrm{d}x \\
&= \int_{-\pi}^{\pi} \frac{e^{i(n+m)x} + e^{i(n-m)x} + e^{-i(n-m)x} + e^{-i(n+m)x}}{4} \,\mathrm{d}x \\
&= \begin{cases} \frac{1}{2} \cdot 2\pi = \pi & (n = m) \\ 0 & (n \neq m) \end{cases}
\end{aligned}$$

□

If (ϕ_n) is an orthonormal system of functions on $[a,b]$, then

$$f \sim \sum_{n=1}^{\infty} c_n \phi_n$$

where $c_n = \langle f, \phi_n \rangle$; we call c_n the n-th Fourier coefficient of f relative to (ϕ_n).

Example. In $\mathbb{R}^3$, let

$$\phi_1 = (1,0,0), \quad \phi_2 = (0,1,0), \quad \phi_3 = (0,0,1).$$

Suppose $f = (2,-1,3)$. Then

$$\langle f, \phi_1 \rangle = 2, \quad \langle f, \phi_2 \rangle = -1, \quad \langle f, \phi_3 \rangle = 3.$$

Hence

$$f \sim 2\phi_1 - \phi_2 + 3\phi_3.$$

The following theorems show that the partial sums of the Fourier series of f have a certain minimum property. We shall assume here and in the rest of this chapter that $f \in \mathscr{R}$, although this hypothesis can be weakened.

Proposition 21.28. *Let (ϕ_n) be an orthonormal system of functions on $[a,b]$. Let*

$$s_n(x) = \sum_{k=1}^{n} c_k \phi_k(x)$$

be the n-th partial sum of the Fourier series of f, and let

$$t_n(x) = \sum_{k=1}^{n} \gamma_k \phi_k(x).$$

Then

$$\|f - s_n\| \leq \|f - t_n\|, \tag{21.12}$$

where equality holds if and only if $\gamma_k = c_k$ for $k = 1, \ldots, n$.

That is to say, among all functions t_n, s_n gives the best possible mean square approximation to f.

Proof. We want to show that

$$\langle f - s_n, f - s_n \rangle \leq \langle f - t_n, f - t_n \rangle .$$

Note that

$$\langle f, s_n \rangle = \left\langle f, \sum_{k=1}^{n} c_k \phi_k \right\rangle = \sum_{k=1}^{n} \overline{c_k} \langle f, \phi_k \rangle = \sum_{k=1}^{n} \overline{c_k} c_k = \sum_{k=1}^{n} |c_k|^2$$

$$\langle s_n, s_n \rangle = \left\langle \sum_{k=1}^{n} c_k \phi_k, \sum_{k=1}^{n} c_k \phi_k \right\rangle = \sum_{k=1}^{n} \langle c_k \phi_k, c_k \phi_k \rangle = \sum_{k=1}^{n} |c_k|^2$$

$$\langle f, t_n \rangle = \sum_{k=1}^{n} c_k \overline{\gamma_k}$$

$$\langle t_n, f \rangle = \sum_{k=1}^{n} \gamma_k \overline{c_k}$$

$$\langle t_n, t_n \rangle = \sum_{k=1}^{n} |\gamma_k|^2$$

Hence we rewrite the desired inequality as

$$\iff \langle f, f \rangle - \sum_{k=1}^{n} |c_k|^2 \leq \langle f, f \rangle - \sum_{k=1}^{n} c_k \overline{\gamma_k} - \sum_{k=1}^{n} \gamma_k \overline{c_k} + \sum_{k=1}^{n} |\gamma_k|^2$$

$$\iff \sum_{k=1}^{n} (c_k \overline{c_k} - c_k \overline{\gamma_k} - \gamma_k \overline{c_k} + \gamma_k \overline{\gamma_k}) \geq 0$$

$$\iff \sum_{k=1}^{n} (c_k - \gamma_k)(\overline{c_k} - \overline{\gamma_k}) \geq 0$$

$$\iff \sum_{k=1}^{n} |c_k - \gamma_k|^2 \geq 0$$

which holds true. Then equality holds if and only if $|c_k - \gamma_k| = 0$, i.e.,

$$\gamma_k = c_k \quad (k = 1, \ldots, n).$$

□

Proposition 21.29 (Bessel inequality). *Let (ϕ_n) be an orthonormal system of functions on $[a,b]$, and*

$$f(x) \sim \sum_{n=1}^{\infty} c_n \phi_n(x).$$

Then

$$\sum_{n=1}^{\infty} |c_n|^2 \leq \|f\|. \tag{21.13}$$

In particular, $c_n \to 0$.

Proof. Letting $n \to \infty$ in (72), we obtain (73) □

the case where equality holds is called Parseval's identity

From now on we shall deal only with the trigonometric system. We shall consider functions f that have period 2π, and are Riemann-integrable on $[-\pi,\pi]$ (and hence on every bounded interval). The Fourier series of f is then the series (63) whose coefficients en are given by the integrals (62), and

$$s_N(x) = s_N(f;x) = \sum_{n=-N}^{N} c_n e^{inx}$$

is the N-th partial sum of the Fourier series of f. The inequality (72) now takes the form

In order to obtain an expression for sN that is more manageable than (75) we introduce the ***Dirichlet kernel***

$$D_N(x) := \sum_{n=-N}^{N} e^{inx}.$$

It follows that

$$\begin{aligned} D_N(x) &= \sum_{n=-N}^{N} e^{inx} \\ &= \frac{e^{-iNx}\left[(e^{ix})^{2N+1}-1\right]}{e^{ix}-1} \\ &= \frac{e^{i(N+1)x}-e^{iNx}}{e^{ix}-1} \\ &= \frac{e^{i(N+\frac{1}{2})x}-e^{-i(N+\frac{1}{2})x}}{e^{\frac{ix}{2}}-e^{\frac{-ix}{2}}} \\ &= \frac{\sin\left(N+\frac{1}{2}\right)x}{\sin\frac{1}{2}x} \end{aligned}$$

Then, for some dummy variable t,

$$\begin{aligned} s_N(x) = \sum_{n=-N}^{N} c_n e^{inx} &= \sum_{n=-N}^{N}\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)e^{-int}\,\mathrm{d}t\right]e^{inx} \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\sum_{n=-N}^{N} f(t)e^{in(x-t)}\right]\mathrm{d}t \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)\left[\sum_{n=-N}^{N} e^{in(x-t)}\right]\mathrm{d}t \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)D_N(x-t)\,\mathrm{d}t\,. \end{aligned}$$

Define the *convolution* of f and g as

$$(f*g)(t) := \int_E f(t)g(x-t)\,\mathrm{d}t\,.$$

The periodicity of all functions involved shows that it is immaterial over which interval we

integrate, as long as its length is 2π. This shows that

$$s_N(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)D_N(x-t)\,\mathrm{d}t = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x-t)D_N(t)\,\mathrm{d}t\,.$$

We shall prove just one result about the pointwise convergence of Fourier series. Before that, we require the following result.

Proposition 21.30 (Riemann–Lebesgue lemma)**.** *Let $f \in \mathscr{R}[a,b]$. Then*

$$\lim_{n\to\infty}\int_a^b f(x)\sin nx\,\mathrm{d}x = 0. \tag{21.14}$$

Proof. □

Proposition 21.31 (Pointwise convergence of Fourier series)**.** *Suppose for some $x \in [-\pi,\pi]$ there exists $M > 0$, $\delta > 0$ such that*

$$\forall t \in (-\delta,\delta), \quad |f(x+t) - f(x)| \le M|t|.$$

Then

$$\lim_{N\to\infty} s_N(f;x) = f(x).$$

Proof. Since

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} D_N(x)\,\mathrm{d}x = 1,$$

we can write

$$f(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)D_N(t)\,\mathrm{d}t\,.$$

Then

$$\begin{aligned} s_N(x) - f(x) &= \frac{1}{2\pi}\int_{-\pi}^{\pi}[f(x-t)-f(x)]D_N(t)\,\mathrm{d}t \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}[f(x-t)-f(x)]\frac{\sin\left(N+\frac{1}{2}\right)t}{\sin\frac{1}{2}t}\,\mathrm{d}t \end{aligned}$$

Let $g(t) = \dfrac{f(x-t)-f(x)}{\sin\frac{1}{2}t}$, then

$$s_N(x) - f(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(t)\sin\left(N+\frac{1}{2}\right)t\,\mathrm{d}t\,.$$

By the Riemann–Lebesgue lemma, we are done. □

Corollary 21.32.

Here is another formulation of this corollary:

This is usually called the localisation theorem. It shows that the behaviour of the sequence $(s_N(f;x))$, as far as convergence is concerned, depends only on the values of f in some (arbitrarily small) neighbourhood of x. Two Fourier series may thus have the same behavior in one interval, but may behave in entirely different ways in some other interval. We have here a very striking contrast between Fourier series and power series (Theorem 8.5).

We conclude with two other approximation theorems.

Theorem 21.33. *If f is continuous (with period 2π) and if $\varepsilon > 0$, then there exists a trigonometric polynomial P such that*

$$|P(x) - f(x)| < \varepsilon \quad (x \in \mathbb{R}).$$

Proof. □

Theorem 21.34 (Parseval's theorem)**.** *Suppose f and g are Riemann-integrable functions with period 2π, and*

$$f(x) \sim \sum_{n=-\infty}^{\infty} c_n e^{inx}, \quad g(x) \sim \sum_{n=-\infty}^{\infty} \gamma_n e^{inx}.$$

Then

(i)
$$\lim_{N\to\infty} \|f - s_N(f)\|^2 = 0.$$

(ii)
$$\frac{1}{2\pi}\langle f, g\rangle = \sum_{n=-\infty}^{\infty} c_n \overline{\gamma_n}.$$

(iii)
$$\|f\|^2 = \sum_{n=-\infty}^{\infty} |c_n|^2.$$

Proof.

(i)

(ii)

(iii)

□

21.4 Gamma Function

The *Gamma function* simulates the factorial.

Definition 21.35 (Gamma function)**.** For $0 < x < \infty$, the ***Gamma function*** is defined as

$$\Gamma(x) := \int_0^\infty t^{x-1} e^{-t}\,\mathrm{d}t\,. \tag{21.15}$$

The integral converges for these x. (When $x < 1$, both 0 and ∞ have to be looked at.)

Lemma 21.36.

(i) The functional equation

$$\Gamma(x+1) = x\Gamma(x)$$

holds for $0 < x < \infty$.

(ii) $\Gamma(n+1) = n!$ *for* $n = 1,2,3,\ldots$

(iii) $\log\Gamma$ *is convex on* $(0,\infty)$.

Proof.

(i) Integrate by parts:

$$\begin{aligned}\Gamma(x+1) &= \int_0^\infty t^x e^{-t}\,\mathrm{d}t \\ &= \left[-t^x e^{-t}\right]_0^\infty + \int_0^\infty x t^{x-1} e^{-t}\,\mathrm{d}t \\ &= 0 + x\Gamma(x) = x\Gamma(x).\end{aligned}$$

(ii) We have

$$\Gamma(1) = \int_0^\infty e^{-t}\,\mathrm{d}t = \left[-e^{-t}\right]_0^\infty = 1.$$

Since $\Gamma(1) = 1$, (i) implies (ii) by induction.

(iii) To show that $\log\Gamma(x)$ is convex, we need to show that for all $p,q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$,

$$\log\Gamma\left(\frac{x}{p} + \frac{y}{q}\right) \geq \frac{1}{p}\log\Gamma(x) + \frac{1}{q}\log\Gamma(y).$$

This is equivalent to showing

$$\Gamma\left(\frac{x}{p} + \frac{y}{q}\right) \geq \Gamma(x)^{\frac{1}{p}} + \Gamma(y)^{\frac{1}{q}}.$$

We have

$$\begin{aligned}
\Gamma\left(\frac{x}{p}+\frac{y}{q}\right) &= \int_0^\infty t^{\frac{x}{p}+\frac{y}{q}-1}e^{-t}\,\mathrm{d}t \\
&= \int_0^\infty t^{\frac{x-1}{p}+\frac{y-1}{q}}+e^{-t\left(\frac{1}{p}+\frac{1}{q}\right)}\,\mathrm{d}t \\
&= \int_0^\infty \left(t^{x-1}e^{-t}\right)^{\frac{1}{p}}\left(t^{y-1}e^{-t}\right)^{\frac{1}{q}}\,\mathrm{d}t \\
&\leq \left[\int_0^\infty \left(t^{\frac{x-1}{p}}e^{-\frac{t}{p}}\right)^p \mathrm{d}t\right]^{\frac{1}{p}}\left[\int_0^\infty \left(t^{\frac{y-1}{q}}e^{-\frac{t}{q}}\right)^q \mathrm{d}t\right]^{\frac{1}{q}} \\
&= \Gamma(x)^{\frac{1}{p}}\Gamma(y)^{\frac{1}{q}}
\end{aligned}$$

where the penultimate line holds as a result of Holder's inequality.

□

In fact, these three properties characterise Γ completely.

Lemma 21.37 (Characterisation of Γ). *If f is a positive function on $(0,\infty)$ such that*

(i) $f(x+1) = xf(x)$,

(ii) $f(1) = 1$,

(iii) $\log f$ *is convex,*

then $f(x) = \Gamma(x)$.

Proof. □

Definition 21.38 (Beta function). For $x > 0$ and $y > 0$, the ***beta function*** is defined as

$$B(x,y) := \int_0^1 t^{x-1}(1-t)^{y-1}\,\mathrm{d}t\,.$$

Lemma 21.39.

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

Proof. Let $f(x) = \dfrac{\Gamma(x+y)}{\Gamma(y)}B(x,y)$. We want to prove that $f(x) = \Gamma(x)$, using 21.37.

(i)

$$B(x+1,y) = \int_0^1 t^{x}(1-t)^{y-1}\,\mathrm{d}t\,.$$

Integrating by parts gives

$$\begin{aligned} B(x+1,y) &= \underbrace{\left[t^x \cdot \frac{(1-t)^y}{y}(-1)\right]_0^1}_{0} + \int_0^1 x t^{x-1} \frac{(1-t)^y}{y}\,\mathrm{d}t \\ &= \frac{x}{y}\int_0^1 t^{x-1}(1-t)^{y-1}(1-t)\,\mathrm{d}t \\ &= \frac{x}{y}\left(\int_0^1 t^{x-1}(1-t)^{y-1}\,\mathrm{d}t - \int_0^1 t^x(1-t)^{y-1}\,\mathrm{d}t\right) \\ &= \frac{x}{y}\left(B(x,y) - B(x+1,y)\right) \end{aligned}$$

which gives $B(x+1,y) = \dfrac{x}{x+y}B(x,y)$. Thus

$$\begin{aligned} f(x+1) &= \frac{\Gamma(x+1+y)}{\Gamma(y)}B(x+1,y) \\ &= \frac{(x+y)B(x+y)}{\Gamma(y)} \cdot \frac{x}{x+y}B(x,y) \\ &= xf(x). \end{aligned}$$

(ii)

$$B(1,y) = \int_0^1 (1-t)^{y-1}\,\mathrm{d}t = \left[-\frac{(1-t)^y}{y}\right]_0^1 = \frac{1}{y}$$

and thus

$$f(1) = \frac{\Gamma(1+y)}{\Gamma(y)}B(1,y) = \frac{y\Gamma(y)}{\Gamma(y)}\frac{1}{y} = 1.$$

(iii) We now show that $\log B(x,y)$ is convex, so that

$$\log f(x) = \underbrace{\log\Gamma(x+y)}_{\text{convex}} + \log B(x,y) - \underbrace{\log\Gamma(y)}_{\text{constant}}$$

is convex with respect to x.

$$B(x_1,y)^{\frac{1}{p}}B(x_2,y)^{\frac{1}{q}} = \left(\int_0^1 t^{x_1-1}(1-t)^{y-1}\,\mathrm{d}t\right)^{\frac{1}{p}}\left(\int_0^1 t^{x_2-1}(1-t)^{y-1}\,\mathrm{d}t\right)^{\frac{1}{q}}$$

By Hölder's inequality,

$$\begin{aligned} B(x_1,y)^{\frac{1}{p}}B(x_2,y)^{\frac{1}{q}} &= \int_0^1 \left[t^{x_1-1}(1-t)^{y-1}\right]^{\frac{1}{p}} \left[t^{x_2-1}(1-t)^{y-1}\right]^{\frac{1}{q}} \mathrm{d}t \\ &= \int_0^1 t^{\frac{x_1}{p}+\frac{x_2}{q}-1}(1-t)^{y-1}\,\mathrm{d}t \\ &= B\left(\frac{x_1}{p}+\frac{x_2}{q},y\right). \end{aligned}$$

Taking log on both sides gives

$$\log B(x,y)^{\frac{1}{p}}B(x_2,y)^{\frac{1}{q}} \geq \log B\left(\frac{x_1}{p}+\frac{x_2}{q},y\right)$$

or

$$\frac{1}{p}\log B(x,y)+\frac{1}{q}\log B(x_2,y) \geq \log B\left(\frac{x_1}{p}+\frac{x_2}{q},y\right).$$

Hence $\log B(x,y)$ is convex, so $\log f(x)$ is convex.

Therefore $f(x)=\Gamma(x)$ which implies $B(x,y)=\dfrac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$. □

An alternative form of Γ is as follows:

$$\Gamma(x) = 2\int_0^{+\infty} t^{2x-1}e^{-t^2}\,\mathrm{d}t.$$

Using this form of Γ, we present an alternative proof.

Proof.

$$\begin{aligned} \Gamma(x)\Gamma(y) &= \left(2\int_0^{+\infty} t^{2x-1}e^{-t^2}\,\mathrm{d}t\right)\left(2\int_0^{+\infty} s^{2y-1}e^{-s^2}\,\mathrm{d}s\right) \\ &= 4\iint_{[0,+\infty)\times[0,+\infty)} t^{2x-1}s^{2y-1}e^{-\left(t^2+s^2\right)}\,\mathrm{d}t\,\mathrm{d}s \end{aligned}$$

Using polar coordinates transformation, let $t=r\cos\theta$, $s=r\sin\theta$. Then $\mathrm{d}t\,\mathrm{d}s = r\,\mathrm{d}r\,\mathrm{d}\theta$. Thus

$$\begin{aligned} \Gamma(x)\Gamma(y) &= 4\int_0^{\frac{\pi}{2}}\left[\int_0^{+\infty} r^{2x-1}\cos^{2x-1}\theta\cdot r^{2y-1}\sin^{2y-1}\theta\cdot e^{-r^2}\cdot r\,\mathrm{d}r\right]\mathrm{d}\theta \\ &= \underbrace{2\int_0^{\frac{\pi}{2}}\cos^{2x-1}\theta\sin^{2y-1}\theta\,\mathrm{d}\theta}_{B(x,y)}\cdot\underbrace{2\int_0^{+\infty} r^{2(x+y)-1}e^{-r^2}\,\mathrm{d}r}_{\Gamma(x+y)} \end{aligned}$$

since

$$\begin{aligned} B(x,y) &= \int_0^1 t^{x-1}(1-t)^{y-1}\,\mathrm{d}t \quad t=\cos^2\theta \\ &= \int_{\frac{\pi}{2}}^0 \cos^{2(x-1)}\theta \sin^{2(y-1)}\theta \cdot 2\cos\theta(-\sin\theta)\,\mathrm{d}\theta \\ &= 2\int_0^{\frac{\pi}{2}} \cos^{2x-1}\theta \sin^{2y-1}\theta\,\mathrm{d}\theta\,. \end{aligned}$$

Hence $B(x,y) = \dfrac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$. □

More on polar coordinates:

$$I = \int_{-\infty}^{+\infty} e^{-x^2}\,\mathrm{d}x = \sqrt{\pi}. \tag{21.16}$$

Proof.

$$\begin{aligned} I^2 &= \int_{-\infty}^{+\infty} e^{-x^2}\,\mathrm{d}x \int_{-\infty}^{+\infty} e^{-y^2}\,\mathrm{d}y \\ &= \iint_{\mathbb{R}^2} e^{-\left(x^2+y^2\right)}\,\mathrm{d}x\,\mathrm{d}y \quad x=r\cos\theta, y=r\sin\theta \\ &= \int_0^{2\pi} \underbrace{\int_0^{+\infty} e^{-r^2} r\,\mathrm{d}r}_{\text{constant w.r.t. } \theta}\,\mathrm{d}\theta \quad s=r^2, \mathrm{d}s=2r\,\mathrm{d}r \\ &= 2\pi\int_0^{+\infty} e^{-s}\cdot\frac{1}{2}\,\mathrm{d}s \\ &= 2\pi\left[\frac{1}{2}e^{-s}(-1)\right]_0^{\infty} = \pi \end{aligned}$$

and thus

$$I = \int_{-\infty}^{+\infty} e^{-x^2}\,\mathrm{d}x = \sqrt{\pi}.$$

□

From this, we have

$$\Gamma\left(\frac{1}{2}\right) = 2\int_0^{\infty} e^{-t^2}\,\mathrm{d}t = \sqrt{\pi}.$$

Lemma 21.40.

$$\Gamma(x) = \frac{2^{x-1}}{\sqrt{\pi}}\Gamma\left(\frac{x}{2}\right)\Gamma\left(\frac{x+1}{2}\right).$$

Proof. Let $f(x) = \dfrac{2^{x-1}}{\sqrt{\pi}}\Gamma\left(\dfrac{x}{2}\right)\Gamma\left(\dfrac{x+1}{2}\right)$. We want to prove that $f(x) = \Gamma(x)$.

(i)

$$\begin{aligned} f(x+1) &= \frac{2^x}{\sqrt{\pi}}\Gamma\left(\frac{x+1}{2}\right)\Gamma\left(\frac{x}{2}+1\right) \\ &= \frac{2^x}{\sqrt{\pi}}\Gamma\left(\frac{x+1}{2}\right)\frac{x}{2}\Gamma\left(\frac{x}{2}\right) \\ &= xf(x) \end{aligned}$$

(ii) $f(1) = \frac{1}{\sqrt{\pi}}\Gamma\left(\frac{1}{2}\right)\Gamma(1) = 1$ since $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

(iii)

$$\log f(x) = \underbrace{(x-1)\log 2}_{\text{linear}} + \underbrace{\log\Gamma\left(\frac{x}{2}\right)}_{\text{convex}} + \underbrace{\log\Gamma\left(\frac{x+1}{2}\right)}_{\text{convex}} - \underbrace{\log\sqrt{\pi}}_{\text{constant}}$$

and hence $\log f(x)$ is convex.

Therefore $f(x) = \Gamma(x)$. □

Theorem 21.41 (Stirling's formula). *This provides a simple approximate expression for $\Gamma(x+1)$ when x is large (hence for $n!$ when n is large). The formula is*

$$\lim_{x\to\infty} \frac{\Gamma(x+1)}{(x/e)^x\sqrt{2\pi x}} = 1. \tag{21.17}$$

Proof. □

Lemma 21.42.

$$B(p,1-p) = \Gamma(p)\Gamma(1-p) = \frac{\pi}{\sin p\pi}.$$

Proof. We have

$$\begin{aligned} B(p,1-p) &= \int_0^1 t^{p-1}(1-t)^{-p}\,\mathrm{d}t \\ &= \int_0^\infty \left(\frac{x}{1+x}\right)^{p-1}\left(\frac{1}{1+x}\right)^{-p}\frac{1}{(1+x)^2}\,\mathrm{d}x \quad \left[x = \frac{t}{1-t}\right] \\ &= \int_0^\infty \frac{x^{p-1}}{1+x}\,\mathrm{d}x \\ &= \int_0^1 \frac{x^{p-1}}{1+x}\,\mathrm{d}x + \int_1^\infty \frac{x^{p-1}}{1+x}\,\mathrm{d}x \end{aligned}$$

See that

$$\begin{aligned}\int_1^\infty \frac{x^{p-1}}{1+x}\,\mathrm{d}x &= \int_1^0 \frac{y^{1-p}}{1+\frac{1}{y}}\left(-\frac{1}{y^2}\right)\mathrm{d}y \quad [x=\frac{1}{y}]\\ &= \int_0^1 \frac{y^{-p}}{1+y}\,\mathrm{d}y = \int_0^1 \frac{x^{-p}}{1+x}\,\mathrm{d}x\end{aligned}$$

so

$$\begin{aligned}B(p,1-p) &= \int_0^1 \frac{x^{p-1}+x^{-p}}{1+x}\,\mathrm{d}x\\ &= \lim_{r\to 1^-}\int_0^r (x^{p-1}+x^{-p})\sum_{k=0}^\infty (-1)^k x^k\,\mathrm{d}x\\ &= \lim_{r\to 1^-}\int_0^r \left(\sum_{k=0}^\infty (-1)^k x^{k+p-1}+\sum_{k=0}^\infty (-1)^k x^{k-p}\right)\mathrm{d}x\\ &= \lim_{r\to 1^-}\left[\sum_{k=0}^\infty (-1)^k \frac{x^{k+p}}{k+p}+\sum_{k=0}^\infty (-1)^k \frac{x^{k-p+1}}{k-p+1}\right]_0^r\\ &= \sum_{k=0}^\infty (-1)^k \frac{1}{k+p}+\sum_{k=0}^\infty (-1)^k \frac{1}{k-p+1}\\ &= \frac{1}{p}+\sum_{k=1}^\infty (-1)^k \frac{1}{k+p}+\sum_{k=1}^\infty (-1)^{k-1}\frac{1}{k+p}\\ &= \frac{1}{p}+\sum_{k=1}^\infty \frac{(-1)^k 2p}{p^2-k^2}\end{aligned}$$

□

Exercises

V
Multivariable Analysis

22 Differentiation

We shall now switch to a different topic, namely that of differentiation in several variable calculus. More precisely, we shall be dealing with maps $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$ from one Euclidean space to another, and trying to understand what the derivative of such a map is.

22.1 Basic Definitions

Recall that for $f\colon \mathbb{R} \to \mathbb{R}$, we defined the derivative at x as

$$f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}.$$

This equation certainly makes no sense in the general case of a function $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$, but can reformulated in a way that does. If $A\colon \mathbb{R} \to \mathbb{R}$ is the linear map defined by $A(h) = f'(x)\cdot h$, then we can rewrite

$$\lim_{h\to 0} \frac{f(x+h)-f(a)-Ah}{h} = 0.$$

Thus we reformulate the definition of differentiability as follows:

> $f\colon \mathbb{R} \to \mathbb{R}$ is differentiable at $x \in \mathbb{R}$ if there exists a linear map $A \in \mathscr{L}(\mathbb{R},\mathbb{R})$ such that
>
> $$\lim_{h\to 0} \frac{f(x+h)-f(a)-Ah}{h} = 0.$$

In this form, the definition has a simple generalisation to higher dimensions:

Definition 22.1 (Derivative)**.** Let $U \subset \mathbb{R}^n$ be open, $\mathbf{f}\colon U \to \mathbb{R}^m$. We say $\mathbf{f}$ is ***differentiable*** at $\mathbf{x} \in U$ if there exists $A \in \mathscr{L}(\mathbb{R}^n,\mathbb{R}^m)$ such that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})-A\mathbf{h}\|}{\|\mathbf{h}\|} = 0.$$

The linear map A is denoted as $\mathbf{f}'(\mathbf{x})$, and called the ***derivative*** of $\mathbf{f}$ at $\mathbf{x}$.

Remark. $\mathbf{h}$ is a point of $\mathbb{R}^n$, and $\mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})-A\mathbf{h}$ is a point of $\mathbb{R}^m$, so the norm signs are essential.

Remark. If $n=m=1$, then $\mathbf{f}'(\mathbf{x})$ coincides with the familiar $f'(x)$ in single-variable calculus.

If $\mathbf{f}$ is differentiable at every $\mathbf{x}\in U$, we say $\mathbf{f}$ is *differentiable on* U.

Remark. Suppose $\mathbf{f}$ is differentiable on U. For every $\mathbf{x}\in U$, $\mathbf{f}'(\mathbf{x})$ is then a function, namely, a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$.

But $\mathbf{f}'$ is also a function: $\mathbf{f}'$ maps U into $\mathscr{L}(\mathbb{R}^n,\mathbb{R}^m)$.

The justification for the phrase "*the* linear map" is as follows:

Lemma 22.2 (Uniqueness of derivative). *Let $U\subset\mathbb{R}^n$ be open, $\mathbf{f}\colon U\to\mathbb{R}^m$. Suppose $\mathbf{x}\in U$, and there exist $A,B\in\mathscr{L}(\mathbb{R}^n,\mathbb{R}^m)$ such that*

$$\lim_{\mathbf{h}\to\mathbf{0}}\frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})-A\mathbf{h}\|}{\|\mathbf{h}\|}=0 \quad \textit{and} \quad \lim_{\mathbf{h}\to\mathbf{0}}\frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})-B\mathbf{h}\|}{\|\mathbf{h}\|}=0.$$

Then $A=B$.

Proof. Suppose $\mathbf{h}\neq\mathbf{0}$. Then

$$\begin{aligned}\lim_{\mathbf{h}\to\mathbf{0}}\frac{\|(A-B)\mathbf{h}\|}{\|\mathbf{h}\|} &= \lim_{\mathbf{h}\to\mathbf{0}}\frac{\|(\mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})-A\mathbf{h})-(\mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})-B\mathbf{h})\|}{\|\mathbf{h}\|}\\ &\le \lim_{\mathbf{h}\to\mathbf{0}}\frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})-A\mathbf{h}\|}{\|\mathbf{h}\|}+\lim_{\mathbf{h}\to\mathbf{0}}\frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})-B\mathbf{h}\|}{\|\mathbf{h}\|}\\ &= 0+0=0.\end{aligned}$$

If $\mathbf{x}\in\mathbb{R}^n$, then $t\mathbf{x}\to 0$ as $t\to 0$. Hence for $\mathbf{x}\neq\mathbf{0}$ we have

$$0=\lim_{t\to 0}\frac{\|A(t\mathbf{x})-B(t\mathbf{x})\|}{\|t\mathbf{x}\|}=\frac{\|A\mathbf{x}-B\mathbf{x}\|}{\|\mathbf{x}\|}.$$

Therefore $A\mathbf{x}=B\mathbf{x}$, which implies $A=B$. □

Example. If $\mathbf{f}\colon\mathbb{R}^n\to\mathbb{R}^m$ is a constant function, then $\mathbf{f}'(\mathbf{x})=\mathbf{0}$.

Proof. Suppose $\mathbf{f}(\mathbf{x})=\mathbf{y}$ for all $\mathbf{x}\in\mathbb{R}^n$. Then

$$\lim_{\mathbf{h}\to\mathbf{0}}\frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})-\mathbf{0}\|}{\|\mathbf{h}\|}=\lim_{\mathbf{h}\to\mathbf{0}}\frac{\|\mathbf{y}-\mathbf{y}-\mathbf{0}\|}{\|\mathbf{h}\|}=0.$$

□

Example. If $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$ is a linear map and $\mathbf{x} \in \mathbb{R}^n$, then $\mathbf{f}'(\mathbf{x}) = \mathbf{f}$.

Remark. Note that $\mathbf{x}$ appears on the LHS, but not on the RHS; the terms on both sides are members of $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$, whereas $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^m$.

Proof. By the linearity of $\mathbf{f}$, we have

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{h})\|}{\|\mathbf{h}\|} = \lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x}) + \mathbf{f}(\mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{h})\|}{\|\mathbf{h}\|} = 0.$$

□

22.2 Basic Theorems

Lemma 22.3 (Differentiability implies continuity). *Let $U \subset \mathbb{R}^n$ be open, suppose $\mathbf{f}\colon U \to \mathbb{R}^m$ is differentiable at $\mathbf{x} \in U$. Then $\mathbf{f}$ is continuous at $\mathbf{x}$.*

Proof. Another way to write the differentiability of $\mathbf{f}$ at $\mathbf{x}$ is to consider the *remainder*:

$$\mathbf{r}(\mathbf{h}) := \mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}.$$

By definition, $\mathbf{f}$ is differentiable at $\mathbf{x}$ if $\lim_{\mathbf{h}\to\mathbf{0}} \|\mathbf{r}(\mathbf{h})\|/\|\mathbf{h}\| = 0$. Thus $\lim_{\mathbf{h}\to\mathbf{0}} \mathbf{r}(\mathbf{h}) = \mathbf{0}$.

The mapping $\mathbf{h} \mapsto \mathbf{f}'(\mathbf{x})\mathbf{h}$ is a linear mapping between finite-dimensional spaces, hence continuous and $\lim_{\mathbf{h}\to\mathbf{0}} \mathbf{f}'(\mathbf{x})\mathbf{h} = \mathbf{0}$. Hence

$$\lim_{\mathbf{h}\to\mathbf{0}} \mathbf{f}(\mathbf{x}+\mathbf{h}) = \mathbf{f}(\mathbf{x}).$$

This precisely means that $\mathbf{f}$ is continuous at $\mathbf{x}$. □

The next result implies that differentiation is a linear map on the space of differentiable functions.

Lemma 22.4. *Let $U \subset \mathbb{R}^n$ be open, suppose $\mathbf{f}, \mathbf{g}\colon U \to \mathbb{R}^m$ are differentiable at $\mathbf{x} \in U$, let $\alpha \in \mathbb{R}$. Then*

(i) $\mathbf{f}+\mathbf{g}$ is differentiable at $\mathbf{x}$, and *(addition)*

$$(\mathbf{f}+\mathbf{g})'(\mathbf{x}) = \mathbf{f}'(\mathbf{x}) + \mathbf{g}'(\mathbf{x}).$$

(ii) $\alpha\mathbf{f}$ is differentiable at $\mathbf{x}$, and *(scalar multiplication)*

$$(\alpha\mathbf{f})'(\mathbf{x}) = \alpha\mathbf{f}'(\mathbf{x}).$$

Proof. Let $\mathbf{h} \in \mathbb{R}^n$, $\mathbf{h} \neq \mathbf{0}$.

(i) We have

$$\begin{aligned}&\frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) + \mathbf{g}(\mathbf{x}+\mathbf{h}) - (\mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x})) - (\mathbf{f}'(\mathbf{x}) + \mathbf{g}'(\mathbf{x}))\mathbf{h}\|}{\|\mathbf{h}\|} \\ &\leq \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}\|}{\|\mathbf{h}\|} + \frac{\|\mathbf{g}(\mathbf{x}+\mathbf{h}) - \mathbf{g}(\mathbf{x}) - \mathbf{g}'(\mathbf{x})\mathbf{h}\|}{\|\mathbf{h}\|}\end{aligned}$$

Then take limits $\mathbf{h} \to \mathbf{0}$ on both sides of the equation.

(ii) Write

$$\frac{\|\alpha\mathbf{f}(\mathbf{x}+\mathbf{h}) - \alpha\mathbf{f}(\mathbf{x}) - \alpha\mathbf{f}'(\mathbf{x})\mathbf{h}\|}{\|\mathbf{h}\|} = |\alpha| \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}\|}{\|\mathbf{h}\|}.$$

Then take limits $\mathbf{h} \to \mathbf{0}$ on both sides of the equation.

□

We now extend the chain rule to the present situation.

Lemma 22.5 (Chain rule). *Let $U \subset \mathbb{R}^n$, $V \subset \mathbb{R}^m$ be open. Suppose $\mathbf{f}\colon U \to \mathbb{R}^m$ is differentiable at $\mathbf{x} \in U$, $\mathbf{f}(U) \subset V$, and $g : V \to \mathbb{R}^k$ is differentiable at $\mathbf{f}(\mathbf{x})$.*
Then $\mathbf{g} \circ \mathbf{f}$ is differentiable at $\mathbf{x}$, and

$$(\mathbf{g} \circ \mathbf{f})'(\mathbf{x}) = \mathbf{g}'(\mathbf{f}(\mathbf{x}))\,\mathbf{f}'(\mathbf{x}).$$

Proof. Let $\mathbf{F} = \mathbf{g} \circ \mathbf{f}$. Let $A = \mathbf{f}'(\mathbf{x})$ and $B = \mathbf{g}'(\mathbf{f}(\mathbf{x}))$. We will show that $\mathbf{f}'(\mathbf{x}) = BA$.

Take a non-zero $\mathbf{h} \in \mathbb{R}^n$ and write $\mathbf{y} = \mathbf{f}(\mathbf{x})$, $\mathbf{k} = \mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x})$. Let

$$\mathbf{r}(\mathbf{h}) = \mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}.$$

Then $\mathbf{r}(\mathbf{h}) = \mathbf{k} - A\mathbf{h}$ or $A\mathbf{h} = \mathbf{k} - \mathbf{r}(\mathbf{h})$, and $\mathbf{f}(\mathbf{x}+\mathbf{h}) = \mathbf{y} + \mathbf{k}$. We look at the quantity we need to go to zero:

$$\begin{aligned}
\frac{\|\mathbf{F}(\mathbf{x}+\mathbf{h}) - \mathbf{F}(\mathbf{x}) - BA\mathbf{h}\|}{\|\mathbf{h}\|} &= \frac{\|\mathbf{g}(\mathbf{f}(\mathbf{x}+\mathbf{h})) - \mathbf{g}(\mathbf{f}(\mathbf{x})) - BA\mathbf{h}\|}{\|\mathbf{h}\|} \\
&= \frac{\|\mathbf{g}(\mathbf{y}+\mathbf{k}) - \mathbf{g}(\mathbf{y}) - B(\mathbf{k} - \mathbf{r}(\mathbf{h}))\|}{\|\mathbf{h}\|} \\
&\le \frac{\|\mathbf{g}(\mathbf{y}+\mathbf{k}) - \mathbf{g}(\mathbf{y}) - B\mathbf{k}\|}{\|\mathbf{h}\|} + \|B\|\frac{\|\mathbf{r}(\mathbf{h})\|}{\|\mathbf{h}\|} \\
&= \frac{\|\mathbf{g}(\mathbf{y}+\mathbf{k}) - \mathbf{g}(\mathbf{y}) - B\mathbf{k}\|}{\|\mathbf{k}\|}\frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x})\|}{\|\mathbf{h}\|} + \|B\|\frac{\|\mathbf{r}(\mathbf{h})\|}{\|\mathbf{h}\|}.
\end{aligned}$$

Take the limit $\mathbf{h} \to \mathbf{0}$. We examine the three terms:

- Since $\mathbf{f}$ is differentiable at $\mathbf{x}$, $\lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{\|\mathbf{r}(\mathbf{h})\|}{\|\mathbf{h}\|} = 0$.
- Since $\mathbf{f}$ is continuous at $\mathbf{x}$, $\mathbf{k} \to \mathbf{0}$ as $\mathbf{h} \to \mathbf{0}$. Thus since $\mathbf{g}$ is differentiable at $\mathbf{y}$,

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{g}(\mathbf{y}+\mathbf{k}) - \mathbf{g}(\mathbf{y}) - B\mathbf{k}\|}{\|\mathbf{k}\|} = 0.$$

- We have

$$\begin{aligned}
\frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x})\|}{\|\mathbf{h}\|} &\le \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}\|}{\|\mathbf{h}\|} + \frac{\|A\mathbf{h}\|}{\|\mathbf{h}\|} \\
&\le \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}\|}{\|\mathbf{h}\|} + \|A\|.
\end{aligned}$$

Since $\mathbf{f}$ is differentiable at $\mathbf{x}$, for small enough $\mathbf{h}$, the quantity $\frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})-A\mathbf{h}\|}{\|\mathbf{h}\|}$ is bounded. Thus the term $\frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})\|}{\|\mathbf{h}\|}$ stays bounded as $\mathbf{h} \to \mathbf{0}$.

Therefore

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{F}(\mathbf{x}+\mathbf{h}) - \mathbf{F}(\mathbf{x}) - BA\mathbf{h}\|}{\|\mathbf{h}\|} = 0,$$

so $\mathbf{f}'(\mathbf{x}) = BA$ as desired. □

We are now assured of the differentiability of those functions $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$, whose component functions are obtained by addition, multiplication, division, and composition, from the functions π_i (which are linear maps) and the functions which we can already differentiate by elementary calculus.

Finding $\mathbf{f}'(\mathbf{x})$, however, may be a fairly formidable task.

Example. Let $f\colon \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(x,y) = \sin\left(xy^2\right)$. Since $f = \sin\circ(\pi_1 \cdot (\pi_2)^2)$, we have

$$\begin{aligned} f'(x,y) &= \sin'(xy^2)\cdot[y^2(\pi_1)'(x,y) + x((\pi_2)^2)'(x,y)] \\ &= \sin'(xy^2)\cdot[y^2(\pi_1)'(x,y) + 2xy(\pi_2)'(x,y)] \\ &= (\cos\left(xy^2\right))\cdot[y^2(1,0) + 2xy(0,1)] \\ &= \left(y^2\cos\left(xy^2\right), 2xy\cos\left(xy^2\right)\right). \end{aligned}$$

Fortunately, we will soon discover a much simpler method of computing $\mathbf{f}'$.

22.3 Partial Derivatives

We begin the attack on the problem of finding derivatives "one variable at a time"; that is, we hold all but one variables constant and take the regular derivative. This is known as a *partial derivative*.

Definition 22.6 (Partial derivative)**.** Let $U \subset \mathbb{R}^n$ be open, $\mathbf{f}\colon U \to \mathbb{R}^m$. The j-th partial derivative at $\mathbf{x} = (x_1, \ldots, x_n) \in U$ is

$$\begin{aligned} D_j\mathbf{f}(\mathbf{x}) &= \lim_{h\to 0} \frac{\mathbf{f}(x_1, \ldots, x_j + h, \ldots, x_n) - \mathbf{f}(x_1, \ldots, x_n)}{h} \\ &= \lim_{t\to 0} \frac{\mathbf{f}(\mathbf{x} + t\mathbf{e}_j) - \mathbf{f}(\mathbf{x})}{t} \quad (j = 1, \ldots, n) \end{aligned}$$

provided the limit exists.

Notation. Some authors also denote the j-th partial derivative at $\mathbf{x}$ by $\dfrac{\partial \mathbf{f}}{\partial x_j}(\mathbf{x})$.

Note that $D_j\mathbf{f}(\mathbf{x})$ is the ordinary derivative of a certain function; if $g\colon \mathbb{R} \to \mathbb{R}^m$ is defined by

$$\mathbf{g}(x) = \mathbf{f}(x_1, \ldots, x, \ldots, x_n),$$

then $D_j\mathbf{f}(\mathbf{x}) = g'(x_j)$. This implies that the computation of $D_j\mathbf{f}(\mathbf{x})$ is a problem we can already solve. If $\mathbf{f}(x_1, \ldots, x_n)$ is given by some formula involving $x_1, \ldots, x_n$, then we find $D_j\mathbf{f}(x_1, \ldots, x_n)$ by differentiating the function whose value at x_j is given by the formula when all x_i, $i \neq j$, are thought of as constants.

Example. If $f(x,y) = \sin\left(xy^2\right)$, then

$$\begin{aligned} D_1 f(x,y) &= \frac{\partial f}{\partial x} = y^2 \cos\left(xy^2\right), \\ D_2 f(x,y) &= \frac{\partial f}{\partial y} = 2xy \cos\left(xy^2\right). \end{aligned}$$

If $f(x,y) = x^y$, then

$$\begin{aligned} D_1 f(x,y) &= \frac{\partial f}{\partial x} = yx^{y-1}, \\ D_2 f(x,y) &= \frac{\partial f}{\partial y} = x^y \log x. \end{aligned}$$

Partial derivatives can be used to find the maxima and minima of functions.

Proposition 22.7. *Let $U \subset \mathbb{R}^n$ be open. If the maximum (or minimum) of $\mathbf{f}\colon U \to \mathbb{R}$ is at $\mathbf{x} \in U$, then*

$$D_j\mathbf{f}(\mathbf{x}) = 0,$$

provided the partial derivatives exists.

Proof. For each $j = 1,\dots,n$, define $g_j(x)\colon \mathbb{R} \to \mathbb{R}^m$ by

$$g_j(\mathbf{x}) = \mathbf{f}(x_1,\dots,x,\dots,x_n).$$

Then $g_j'(x_j) = D_j\mathbf{f}(\mathbf{x})$.

Clearly g_j has a maximum (or minimum) at x_j, and g_j is defined in an open interval containing x_j. Hence $D_j\mathbf{f}(\mathbf{x}) = g_j'(x_j) = 0$. □

Remark. The converse is false even if $n = 1$ (if $f\colon \mathbb{R} \to \mathbb{R}$ is defined by $f(x) = x^3$, then $f'(0) = 0$, but 0 is not even a local maximum or minimum).

If $n > 1$, the converse may fail to be true in a rather spectacular way. Let $f\colon \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(x,y) = x^2 - y^2$. Since g_1 has a minimum at 0, and g_2 has a maximum at 0,

$$D_1 f(0,0) = 0,$$
$$D_2 f(0,0) = 0,$$

but clearly $(0,0)$ is neither a local maximum or minimum.

Let $\{\mathbf{e}_1,\dots,\mathbf{e}_n\}$ and $\{\mathbf{u}_1,\dots,\mathbf{u}_m\}$ be the standard bases of $\mathbb{R}^n$ and $\mathbb{R}^m$. Let $U \subset \mathbb{R}^n$ be open, $\mathbf{f}\colon U \to \mathbb{R}^m$. The *components* of $\mathbf{f}$ are the real-valued functions $f_1,\dots,f_m\colon U \to \mathbb{R}$ defined by

$$\mathbf{f}(\mathbf{x}) = \sum_{i=1}^{m} f_i(\mathbf{x})\mathbf{u}_i \quad (\mathbf{x} \in U).$$

That is, $\mathbf{f}(\mathbf{x})$ is the point in $\mathbb{R}^m$ with coordinates

$$\mathbf{f}(\mathbf{x}) = \left(f_1(\mathbf{x}),\dots,f_m(\mathbf{x})\right).$$

The next result states that in order to differentiate a function $\mathbf{f}$, we can differentiate each of its components.

Lemma 22.8. *If $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$, then $\mathbf{f}$ is differentiable at $\mathbf{x} \in \mathbb{R}^n$ if and only if each component f_i is differentiable at $\mathbf{x}$, and*

$$\mathbf{f}'(\mathbf{x}) = \left(f_1'(\mathbf{x}),\dots,f_m'(\mathbf{x})\right).$$

Thus $\mathbf{f}'(\mathbf{x})$ is the $m \times n$ matrix whose i-th row is $(f_i)'(\mathbf{x})$.

Proof. Suppose each f_i is differentiable at $\mathbf{x}$. Let

$$A = \left(f_1'(\mathbf{x}), \ldots, f_m'(\mathbf{x})\right).$$

Then

$$\begin{aligned}
&\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h} \\
&= (f_1(\mathbf{x}+\mathbf{h}), \ldots, f_m(\mathbf{x}+\mathbf{h})) - (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})) - \left(f_1'(\mathbf{x})\mathbf{h}, \ldots, f_m'(\mathbf{x})\mathbf{h}\right) \\
&= \left(f_1(\mathbf{x}+\mathbf{h}) - f_1(\mathbf{x}) - f_1'(\mathbf{x})\mathbf{h}, \ldots, f_m(\mathbf{x}+\mathbf{h}) - f_m(\mathbf{x}) - f_m'(\mathbf{x})\mathbf{h}\right).
\end{aligned}$$

Therefore

$$\lim_{h\to 0} \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}\|}{\|\mathbf{h}\|} \leq \lim_{h\to 0} \sum_{i=1}^{m} \frac{\|f_i(\mathbf{x}+\mathbf{h}) - f_i(\mathbf{x}) - f_i'(\mathbf{x})\mathbf{h}\|}{\|\mathbf{h}\|} = 0.$$

□

22.4 Derivatives

Likewise we can also take partial derivatives of each component: the j-th partial derivative of the i-th component of $\mathbf{f}$ at $\mathbf{x} \in U$ is

$$D_j f_i(\mathbf{x}) := \lim_{t \to 0} \frac{f_i(\mathbf{x} + t\mathbf{e}_j) - f_i(\mathbf{x})}{t} \quad (1 \le i \le m,\ 1 \le j \le n),$$

provided the limit exists.

Notation. Writing $f_i(x_1, \ldots, x_n)$ in place of $f_i(\mathbf{x})$, we see that $D_j f_i$ is the derivative of f_i with respect to x_j, keeping the other variables fixed. Hence the notation $\frac{\partial f_i}{\partial x_j}$ is often used in place of $D_j f_i$.

The next result states that if $\mathbf{f}$ is differentiable at a point $\mathbf{x}$, then its partial derivatives exist at $\mathbf{x}$, and they determine the linear map $\mathbf{f}'(\mathbf{x})$ completely:

Theorem 22.9. *Let $U \subset \mathbb{R}^n$ be open, suppose $\mathbf{f}\colon U \to \mathbb{R}^m$ is differentiable at $\mathbf{x} \in U$. Then all the partial derivatives at $\mathbf{x}$ exist, and the matrix of $\mathbf{f}'(\mathbf{x})$ with respect to the standard bases of $\mathbb{R}^n$ and $\mathbb{R}^m$ is*

$$\left[\mathbf{f}'(\mathbf{x})\right]_{ij} = D_j f_i(\mathbf{x}).$$

That is,

$$\left[\mathbf{f}'(\mathbf{x})\right] = \begin{pmatrix} D_1 f_1(\mathbf{x}) & D_2 f_1(\mathbf{x}) & \cdots & D_n f_1(\mathbf{x}) \\ D_1 f_2(\mathbf{x}) & D_2 f_2(\mathbf{x}) & \cdots & D_n f_2(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ D_1 f_m(\mathbf{x}) & D_2 f_m(\mathbf{x}) & \cdots & D_n f_m(\mathbf{x}) \end{pmatrix}.$$

We call $[\mathbf{f}'(\mathbf{x})]$ the ***Jacobian matrix*** of $\mathbf{f}$ at $\mathbf{x}$.

Remark. If $f\colon \mathbb{R} \to \mathbb{R}$, then $f'(x)$ is a 1×1 matrix whose single entry is the number denoted as $f'(x)$ in single-variable calculus.

Proof. Fix $j \in \{1, \ldots, n\}$. We want to show that

$$\mathbf{f}'(\mathbf{x})\mathbf{e}_j = \sum_{i=1}^{m} D_j f_i(\mathbf{x})\mathbf{u}_i.$$

Since $\mathbf{f}$ is differentiable at $\mathbf{x}$, writing differentiability in terms of the remainder gives us

$$\mathbf{f}(\mathbf{x} + t\mathbf{e}_j) - \mathbf{f}(\mathbf{x}) = \mathbf{f}'(\mathbf{x})(t\mathbf{e}_j) + \mathbf{r}(t\mathbf{e}_j)$$

where $\left\|\mathbf{r}(t\mathbf{e}_j)\right\|/t \to 0$ as $t \to 0$. Taking the limit $t \to 0$ on both sides, the linearity of $\mathbf{f}'(\mathbf{x})$ shows that

$$\lim_{t \to 0} \frac{\mathbf{f}(\mathbf{x} + t\mathbf{e}_j) - \mathbf{f}(\mathbf{x})}{t} = \mathbf{f}'(\mathbf{x})\mathbf{e}_j.$$

If we now represent $\mathbf{f}$ in terms of its components, the above equation becomes

$$\lim_{t\to 0}\sum_{i=1}^{m}\frac{f_i(\mathbf{x}+t\mathbf{e}_j)-f_i(\mathbf{x})}{t}\mathbf{u}_i=\mathbf{f}'(\mathbf{x})\mathbf{e}_j.$$

It follows that each quotient in this sum has a limit as $t\to 0$ (see Theorem 4.10), so that each partial derivative D_jf_i exists. Hence

$$\mathbf{f}'(\mathbf{x})\mathbf{e}_j=\sum_{i=1}^{m}D_jf_i(\mathbf{x})\mathbf{u}_i\quad(j=1,\dots,n).$$

□

The converse is true if one hypothesis is added.

Definition 22.10. Let $U\subset\mathbb{R}^n$ be open. We say $\mathbf{f}\colon U\to\mathbb{R}^m$ is ***continuously differentiable*** if $\mathbf{f}$ is differentiable, and $\mathbf{f}'$ is continuous.

We also say that $\mathbf{f}$ is a $\mathscr{C}'$-mapping, or that $\mathbf{f}\in\mathscr{C}'(U)$.

More explicitly, it is required that

$$\forall\varepsilon>0,\quad\exists\delta>0,\quad\forall x\in U,\quad\|\mathbf{x}-\mathbf{y}\|<\delta\implies\left\|\mathbf{f}'(\mathbf{y})-\mathbf{f}'(\mathbf{x})\right\|<\varepsilon.$$

Theorem 22.11. *Let $U\subset\mathbb{R}^n$ be open, $\mathbf{f}\colon U\to\mathbb{R}^m$. Then $\mathbf{f}$ is continuously differentiable if and only if all the partial derivatives D_jf_i exist and are continuous on U.*

Proof.

$\boxed{\implies}$ Suppose $\mathbf{f}\in\mathscr{C}'(U)$. Then $\mathbf{f}$ is differentiable on U, so

$$(D_jf_i)(\mathbf{x})=(\mathbf{f}'(x)\mathbf{e}_j)\cdot\mathbf{u}_i$$

for all i, j, and for all $\mathbf{x}\in U$. Hence

$$(D_jf_i)(\mathbf{y})-(D_jf_i)(\mathbf{x})=\big((\mathbf{f}'(\mathbf{y})-\mathbf{f}'(\mathbf{x}))\mathbf{e}_j\big)\cdot\mathbf{u}_i.$$

Since $\|\mathbf{u}_i\|=\left\|\mathbf{e}_j\right\|=1$, it follows that

$$\begin{aligned}\left\|(D_jf_i)(\mathbf{y})-(D_jf_i)(\mathbf{x})\right\|&\le\left\|(\mathbf{f}'(\mathbf{y})-\mathbf{f}'(\mathbf{x}))\mathbf{e}_j\right\|\\&\le\left\|\mathbf{f}'(\mathbf{y})-\mathbf{f}'(\mathbf{x})\right\|.\end{aligned}$$

Hence D_jf_i is continuous.

$\boxed{\impliedby}$ □

Gradients, Curves, and Directional Derivatives

Let γ be a differentiable mapping of $(a,b) \subset \mathbb{R}$ into an open set $U \subset \mathbb{R}^n$; that is, γ is a differentiable curve in U. Let $f: U \to \mathbb{R}$ be differentiable.

For $t \in (a,b)$, define

$$g(t) = f(\gamma(t)).$$

By the chain rule,

$$g'(t) = f'(\gamma(t))\,\gamma'(t).$$

Since $\gamma'(t) \in \mathscr{L}(\mathbb{R}, \mathbb{R}^n)$ and $f'(\gamma(t)) \in \mathscr{L}(\mathbb{R}^n, \mathbb{R})$, $g'(t)$ is a linear operator on $\mathbb{R}$; thus, we can regard $g'(t)$ as a real number. This number can be computed in terms of the partial derivatives of f and the derivatives of the components of γ_i, as we shall now see.

With respect to the standard basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ of $\mathbb{R}^n$, the matrix of $\gamma'(t)$ is the $n \times 1$ matrix which has $\gamma_i'(t)$ in the i-th row, where $\gamma_1, \ldots, \gamma_n$ are the components of γ. For every $\mathbf{x} \in U$, the matrix of $f'(\mathbf{x})$ is the $1 \times n$ matrix which has $\frac{\partial f}{\partial x_j}$ in the j-th column. Hence the matrix of $g'(t)$ is the 1×1 matrix whose only entry is the real number

$$g'(t) = f'(\gamma(t))\gamma'(t) = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\gamma(t)) \frac{\mathrm{d}\gamma_j}{\mathrm{d}t} = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j} \frac{\mathrm{d}\gamma_j}{\mathrm{d}t}.$$

Definition 22.12 (Gradient). Let $U \subset \mathbb{R}^n$ be open, suppose $f: U \to \mathbb{R}$ is differentiable. The ***gradient*** at $\mathbf{x} \in U$ is defined as

$$(\nabla f)(\mathbf{x}) := \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\mathbf{x}) \mathbf{e}_j. \tag{22.1}$$

Writing $\gamma'(t)$ as components

$$\gamma'(t) = \sum_{j=1}^{n} \gamma_i'(t) \mathbf{e}_j,$$

using the scalar product, we can rewrite $g'(t)$ as

$$g'(t) = (\nabla f)(\gamma(t)) \cdot \gamma'(t). \tag{22.2}$$

Let us now fix $\mathbf{x} \in U$, take a unit vector $\mathbf{u} \in \mathbb{R}^n$, and let γ be

$$\gamma(t) = \mathbf{x} + t\mathbf{u}.$$

Then $\gamma'(t) = \mathbf{u}$ for every t. Hence (22.2) shows that

$$g'(0) = (\nabla f)(\mathbf{x}) \cdot \mathbf{u}.$$

On the other hand, we have

$$g(t) - g(0) = f(\mathbf{x} + t\mathbf{u}) - f(\mathbf{x}).$$

Hence

$$\lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{u}) - f(x)}{t} = (\nabla f)(\mathbf{x}) \cdot \mathbf{u}. \tag{22.3}$$

We call this limit the ***directional derivative*** of f at $\mathbf{x}$, in the direction of the unit vector $\mathbf{u}$, and may be denoted by $(D_{\mathbf{u}} f)(\mathbf{x})$.

If f and $\mathbf{x}$ are fixed, but $\mathbf{u}$ varies, then (22.3) shows that $(D_{\mathbf{u}} f)(\mathbf{x})$ attains it maximum when $\mathbf{u}$ is a positive scalar multiple of $(\nabla f)(\mathbf{x})$. [The case $(\nabla f)(\mathbf{x}) = \mathbf{0}$ should be excluded here.]

If $\mathbf{u} = \sum_j u_j \mathbf{e}_j$, then (22.3) shows that $(D_{\mathbf{u}} f)(\mathbf{x})$ can be expressed in terms of the partial derivatives of f at $\mathbf{x}$:

$$(D_{\mathbf{u}} f)(\mathbf{x}) = \sum_{j=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{x}) \mathbf{u}_i. \tag{22.4}$$

Proposition 22.13. *Let $U \subset \mathbb{R}^n$ be open and convex, suppose $\mathbf{f}: U \to \mathbb{R}^m$ is differentiable on U, and there exists a real number M such that*

$$\left\| \mathbf{f}'(\mathbf{x}) \right\| \leq M \quad (\mathbf{x} \in U).$$

Then for all $\mathbf{a}, \mathbf{b} \in U$,

$$\|\mathbf{f}(\mathbf{b}) - \mathbf{f}(\mathbf{a})\| \leq M \|\mathbf{b} - \mathbf{a}\|.$$

Proof. Fix $\mathbf{a}, \mathbf{b} \in U$. Define

$$\gamma(t) = (1-t)\mathbf{a} + t\mathbf{b}$$

for all $t \in \mathbb{R}$ such that $\gamma(t) \in U$. Since U is convex, $\gamma(t) \in U$ if $0 \leq t \leq 1$. Put

$$\mathbf{g}(t) = \mathbf{f}(\gamma(t)).$$

Then

$$\mathbf{g}'(t) = \mathbf{f}'(\gamma(t))\, \gamma'(t) = \mathbf{f}'(\gamma(t))\, (\mathbf{b} - \mathbf{a}),$$

so that

$$|\mathbf{g}'(t)| \leq \left\| \mathbf{f}'(\gamma(t)) \right\| |\mathbf{b} - \mathbf{a}| \leq M |\mathbf{b} - \mathbf{a}|$$

for all $t \in [0,1]$. By Theorem 5.19,

$$\|\mathbf{g}(1) - \mathbf{g}(0)\| \leq M \|\mathbf{b} - \mathbf{a}\|.$$

But $\mathbf{g}(0) = \mathbf{f}(\mathbf{a})$ and $\mathbf{g}(1) = \mathbf{f}(\mathbf{b})$. This completes the proof. □

Corollary 22.14. *If, in addition,* $\mathbf{f}'(\mathbf{x}) = \mathbf{0}$ *for all* $\mathbf{x} \in U$*, then* $\mathbf{f}$ *is constant.*

Proof. To prove this, note that the hypotheses of the previous result hold now with $M = 0$. □

The Jacobian

Definition 22.15 (Jacobian). Let $U \subset \mathbb{R}^n$, suppose $\mathbf{f}\colon U \to \mathbb{R}^m$ is differentiable. Define the ***Jacobian*** of $\mathbf{f}$ at $\mathbf{x} \in U$ as

$$J_{\mathbf{f}}(\mathbf{x}) := \det\big[\mathbf{f}'(\mathbf{x})\big].$$

We shall also denote $J_{\mathbf{f}}$ as

$$\frac{\partial(f_1,\ldots,f_n)}{\partial(x_1,\ldots,x_n)}.$$

This last piece of notation may seem somewhat confusing, but it is quite useful when we need to specify the exact variables and function components used, as we will do, for example, in the implicit function theorem.

The Jacobian determinant $J_{\mathbf{f}}$ is a real-valued function, and when $n = 1$ it is simply the derivative. From the chain rule and $\det AB = \det A \det B$, it follows that

$$J_{\mathbf{f}\circ\mathbf{g}}(\mathbf{x}) = J_{\mathbf{f}}(\mathbf{g}(\mathbf{x}))J_{\mathbf{g}}(\mathbf{x}).$$

The determinant of a linear mapping tells us what happens to area/volume under the mapping. Similarly, the Jacobian determinant measures how much a differentiable mapping stretches things locally, and if it flips orientation. In particular, if the Jacobian determinant is non-zero than we would assume that locally the mapping is invertible (and we would be correct as we will later see).

Continuity and The Derivative

Let us prove a "mean value theorem" for vector-valued functions.

Theorem 22.16. *If* $\phi : [a,b] \to \mathbb{R}^n$ *is differentiable on* (a,b) *and continuous on* $[a,b]$*, then there exists* $t \in [a,b]$ *such that*

$$\|\phi(b) - \phi(a)\| \leq (b-a)\big\|\phi'(t)\big\|. \tag{22.5}$$

22.5 Inverse Functions

Suppose $f\colon \mathbb{R} \to \mathbb{R}$ is continuously differentiable in an open set containing a, and $f'(a) \neq 0$. If $f'(a) > 0$, there is an open interval V containing a such that $f'(x) > 0$ for all $x \in V$, and a similar statement holds if $f'(a) < 0$.

Thus f is increasing (or decreasing) on V, and is therefore bijective with an inverse function f^{-1} defined on some open interval W containing $f(a)$. Moreover it is not hard to show that f^{-1} is differentiable, and

$$(f^{-1})'(y) = \frac{1}{f'(f^{-1}(y))} \quad (y \in W).$$

An analogous discussion in higher dimensions is much more involved. The inverse function theorem states, roughly speaking, that a continuously differentiable mapping $\mathbf{f}$ is invertible in a neighbourhood of any point $\mathbf{x}$ at which the linear map $\mathbf{f}'(\mathbf{x})$ is invertible:

Theorem 22.17 (Inverse function theorem). *Let $U \subset \mathbb{R}^n$ be open, suppose $\mathbf{f}\colon U \to \mathbb{R}^n$ is a $\mathscr{C}'$-mapping, and $\mathbf{f}'(\mathbf{a})$ is invertible for some $\mathbf{a} \in U$, and $\mathbf{b} = \mathbf{f}(\mathbf{a})$. Then*

(i) there exist open sets $U, V \subset \mathbb{R}^n$ such that $\mathbf{a} \in U$, $\mathbf{b} \in V$, $\mathbf{f}$ is bijective on U, and $\mathbf{f}(U) = V$;

(ii) if $\mathbf{g}$ is the inverse of $\mathbf{f}$ [which exists, by (i)], defined in V by

$$\mathbf{g}(\mathbf{f}(\mathbf{x})) = \mathbf{x} \quad (\mathbf{x} \in U),$$

then $\mathbf{g} \in \mathscr{C}'(V)$.

Writing the equation $\mathbf{y} = \mathbf{f}(\mathbf{x})$ in component form, we arrive at the following interpretation of the conclusion of the theorem: The system of n equations

$$y_i = f_i(x_1, \ldots, x_n) \quad (i = 1, \ldots, n)$$

can be solved for $x_1, \ldots, x_n$ in terms of $y_1, \ldots, y_n$, if we restrict x and y to small enough neighbourhoods of $\mathbf{a}$ and $\mathbf{b}$; the solutions are unique and continuously differentiable.

Corollary 22.18. *Let $U \subset \mathbb{R}^n$ be open, suppose $\mathbf{f}\colon U \to \mathbb{R}^m$ is continuously differentiable on U. If $\mathbf{f}'(\mathbf{x})$ is invertible for every $\mathbf{x} \in U$, then $f(W)$ is an open subset of $\mathbb{R}^n$ for every open set $W \subset U$.*

22.6 Implicit Functions

If f is a continuously differentiable real function in the plane, then the equation $f(x,y)=0$ can be solved for y in terms of x in a neighbourhood of any point (a,b) at which $f(a,b)=0$ and $\frac{\partial f}{\partial y}\neq 0$. Likewise, one can solve for x in terms of y near (a,b) if $\frac{\partial f}{\partial x}\neq 0$ at (a,b). For a simple example which illustrates the need for assuming $\frac{\partial f}{\partial y}\neq 0$, consider $f(x,y)=x^2+y^2-1$.

The preceding very informal statement is the simplest case (the case $m=n=1$ of Theorem 9.28) of the so-called "implicit function theorem". Its proof makes strong use of the fact that continuously differentiable maps behave locally very much like their derivatives. Accordingly, we first prove Theorem 9.27, the linear version of Theorem 9.28.

Notation. In what follows, the first entry in $(\mathbf{x},\mathbf{y})$ or in a similar symbol will always be a vector in $\mathbb{R}^n$, the second will be a vector in $\mathbb{R}^m$.

Every $A\in\mathscr{L}(\mathbb{R}^{n+m},\mathbb{R}^n)$ can be split into two linear maps A_x and A_y, defined by

$$A_x\mathbf{h}=A(\mathbf{h},\mathbf{0}),\quad A_y\mathbf{k}=A(\mathbf{0},\mathbf{k})$$

for any $\mathbf{h}\in\mathbb{R}^n$, $\mathbf{k}\in\mathbb{R}^m$. Then $A_x\in\mathscr{L}(\mathbb{R}^n)$, $A_y\in\mathscr{L}(\mathbb{R}^m,\mathbb{R}^n)$, and

$$A(\mathbf{h},\mathbf{k})=A_x\mathbf{h}+A_y\mathbf{k}.$$

The linear version of the implicit function theorem is now almost obvious.

Theorem 22.19. *If $A\in\mathscr{L}(\mathbb{R}^{n+m},\mathbb{R}^n)$ and if A_x is invertible, then for every $\mathbf{k}\in\mathbb{R}^m$ there exists a unique $\mathbf{h}\in\mathbb{R}^n$ such that $A(\mathbf{h},\mathbf{k})=0$.*
This $\mathbf{h}$ can be computed from $\mathbf{k}$ by the formula

$$\mathbf{h}=-(A_x)^{-1}A_y\mathbf{k}.$$

The conclusion of Theorem 9.27 is, in other words, that the equation $A(\mathbf{h},\mathbf{k})=0$ can be solved (uniquely) for $\mathbf{h}$ if $\mathbf{k}$ is given, and that the solution $\mathbf{h}$ is a linear function of $\mathbf{k}$. Those who have some acquaintance with linear algebra will recognise this as a very familiar statement about systems of linear equations.

Theorem 22.20. *Let $\mathbf{f}$ be a $\mathscr{C}'$-mapping of an open set $U\subset\mathbb{R}^{n+m}$ into $\mathbb{R}^n$, such that $\mathbf{f}(\mathbf{a}+\mathbf{b})=\mathbf{0}$ for some point $(\mathbf{a},\mathbf{b})\in U$.*
Put $A=\mathbf{f}'(\mathbf{a},\mathbf{b})$ and assume that A_x is invertible. Then there exist open sets $U\subset\mathbb{R}^{n+m}$ and $W\subset\mathbb{R}^m$, with $(\mathbf{a},\mathbf{b})\in U$ and $\mathbf{b}\in W$, having the following property: for every $\mathbf{y}\in W$ there exists a unique $\mathbf{x}$ such that

$$(\mathbf{x},\mathbf{y})\in U,\quad \mathbf{f}(\mathbf{x},\mathbf{y})=0.$$

If this $\mathbf{x}$ *is defined to be* $\mathbf{g}(\mathbf{y})$*, then* $\mathbf{g}$ *is a* $\mathscr{C}'$*-mapping of* W *into* $\mathbb{R}^n$*,* $\mathbf{g}(\mathbf{b}) = \mathbf{a}$*,*

$$\mathbf{f}(\mathbf{g}(\mathbf{y}), \mathbf{y}) = \mathbf{0} \quad (\mathbf{y} \in W),$$

and

$$\mathbf{g}'(\mathbf{b}) = -(A_x)^{-1} A_y.$$

22.7 Derivatives of Higher Order

Definition 22.21. Let $U \subset \mathbb{R}^n$ be open, suppose $f: U \to \mathbb{R}$, with partial derivatives $\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}$. If the functions $\frac{\partial f}{\partial x_j}$ are themselves differentiable, then the *second-order partial derivatives* of f are defined by

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right) \quad (i, j = 1, \ldots, n).$$

If all these functions $\frac{\partial^2 f}{\partial x_i \partial x_j}$ are continuous on U, we say that f is of class $\mathscr{C}''$ in U, or that $f \in \mathscr{C}''(U)$.
$\mathbf{f}: U \subset \mathbb{R}^n \to \mathbb{R}^m$ is said to be of class $\mathscr{C}''$ if each component of $\mathbf{f}$ is of class $\mathscr{C}''$.

It can happen that

22.8 Differentiation of Integrals

Exercises

23 Integration

23.1 Basic Definitions

Recall that a partition P of a closed interval $[a,b]$ is a set of points $t_0,\dots,t_k$ where

$$a = t_0 \le t_1 \le \cdots \le t_k = b.$$

The partition P divides the interval $[a,b]$ into k subintervals $[t_{i-1},t_i]$.

More generally, a *partition* of a rectangle $[a_1,b_1]\times\cdots\times[a_n,b_n]$ is a collection

$$P = (P_1,\dots,P_n),$$

where each P_i is a partition of the interval $[a_i,b_i]$.

Example. Suppose $P_1 = \{t_0,\dots,t_k\}$ is a partition of $[a_1,b_1]$ and $P_2 = \{s_0,\dots,s_l\}$ is a partition of $[a_2,b_2]$. Then the partition $P = (P_1,P_2)$ of $[a_1,b_1]\times[a_2,b_2]$ divides the closed rectangle into $k\cdot l$ subrectangles of the form $[t_{i-1},t_i]\times[s_{j-1},s_j]$.

In general, if P_i divides $[a_i,b_i]$ into N_i subintervals, then $P = (P_1,\dots,P_n)$ divides $[a_1,b_1]\times\cdots\times[a_n,b_n]$ into $N = N_1\cdots N_n$ subrectangles. These subrectangles are called *subrectangles of the partition* P.

Let $A\subset\mathbb{R}^n$ be a rectangle, $f\colon A\to\mathbb{R}$ be a bounded function, and P be a partition of A. For each subrectangle S of the partition, let

$$m_S(f) = \inf_{x\in S} f(x)$$
$$M_S(f) = \sup_{x\in S} f(x)$$

and let $v(S)$ be the volume of $S = [a_1,b_1]\times\cdots\times[a_n,b_n]$, defined by

$$v(S) = (b_1-a_1)\cdots(b_n-a_n).$$

The *lower* and *upper sums* of f for P are defined by

$$L(f,P) = \sum_S m_S(f) \cdot v(S),$$
$$U(f,P) = \sum_S M_S(f) \cdot v(S).$$

Clearly $L(f,P) \leq U(f,P)$.

We say a partition P' *refines* P if each subrectangle of P' is in a subrectangle of P.

Lemma 23.1. *Suppose the partition P' refines P. Then*

$$L(f,P) \leq L(f,P') \quad \text{and} \quad U(f,P') \leq U(f,P).$$

Proof. Each subrectangle S of P is divided into several subrectangles $S_1,\ldots,S_\alpha$ of P', so $v(S) = v(S_1) + \cdots + v(S_\alpha)$. Now $m_S(f) \leq m_{S_i}(f)$, since the values $f(x)$ for $x \in S$ include all values $f(x)$ for $x \in S_i$ (and possibly smaller ones). Thus

$$\begin{aligned} m_S(f) \cdot v(S) &= m_S(f) \cdot v(S_1) + \cdots + m_S(f) \cdot v(S_\alpha) \\ &\leq m_{S_1}(f) \cdot v(S_1) + \cdots + m_{S_\alpha}(f) \cdot v(S_\alpha). \end{aligned}$$

The sum, for all S, of the terms on the LHS is $L(f,P)$, while the sum of all the terms on the RHS is $L(f,P')$. Hence $L(f,P) < L(f,P')$. The proof for upper sums is similar. □

Corollary 23.2. *If P and P' are any two partitions, then $L(f,P') \leq U(f,P)$.*

Proof. Let P'' be a partition which refines both P and P'. Then

$$L(f,P') \leq L(f,P'') \leq U(f,P'') \leq U(f,P).$$

□

Define the *upper* and *lower integrals* of f over A by

$$\overline{\int_A} f = \inf_{P \in \mathscr{P}(A)} U(f,P)$$
$$\underline{\int_A} f = \sup_{P \in \mathscr{P}(A)} L(f,P)$$

where $\mathscr{P}(A)$ denotes the set of all partitions of A.

The previous result implies that the sup of all lower sums for f is less than or equal to the inf of

all upper sums for f; in other words,

$$\underline{\int_A} f \leq \overline{\int_A} f.$$

If the two values coincide, we say f is integrable:

Definition 23.3. We say a bounded function $f: A \to \mathbb{R}$ is ***integrable*** on the rectangle A if

$$\underline{\int_A} f = \overline{\int_A} f.$$

This common number is called the *integral* of f over A, and denoted

$$\int_A f.$$

Often, the notation

$$\int_A f(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n$$

is used. If $f: [a,b] \to \mathbb{R}$, then this coincides with the Riemann integral: $\int_a^b f = \int_{[a,b]} f$.

A simple but useful criterion for integrability is provided by the next result.

Lemma 23.4 (Integrability criterion). *A bounded function $f: A \to \mathbb{R}$ is integrable if and only if*

$$\forall \varepsilon > 0, \quad \exists P, \quad U(f,P) - L(f,P) < \varepsilon.$$

This means we can make the upper and lower sums arbitrarily close.

Proof.

$\boxed{\Longrightarrow}$ Suppose f is integrable. Then

$$\sup_{P \in \mathscr{P}(A)} L(f,P) = \inf_{P \in \mathscr{P}(A)} U(f,P).$$

Thus for any $\varepsilon > 0$, there exists partitions P and P' such that

$$U(f,P) - L(f,P') < \varepsilon.$$

Let P'' be a common refinement of P and P'. Then

$$U(f,P'') - L(f,P'') \leq U(f,P) - L(f,P') < \varepsilon.$$

$\boxed{\Longleftarrow}$ Let $\varepsilon > 0$ be given. Suppose there exists a partition P such that $U(f,P) - L(f,P) < \varepsilon$.

Then it is clear that

$$\sup_{P\in\mathscr{P}(A)} L(f,P) = \inf_{P\in\mathscr{P}(A)} U(f,P).$$

Hence f is integrable. □

In the following sections we will characterize the integrable functions and discover a method of computing integrals. For the present we consider two functions, one integrable and one not.

Example (Constant function). Let $f\colon A\to\mathbb{R}$ be a constant function, $f(x)=c$. Then for any partition P and subrectangle S, we have

$$m_S(f) = M_S(f) = c,$$

so that

$$L(f,P) = U(f,P) = \sum_S c\cdot v(S) = c\cdot v(A).$$

Hence $\int_A f = c\cdot v(A)$.

Example (Dirichlet's function). Let $f\colon [0,1]\times[0,1]\to\mathbb{R}$ be defined by

$$f(x,y) = \begin{cases} 0 & (x\in\mathbb{Q}) \\ 1 & (x\in\mathbb{R}\setminus\mathbb{Q}) \end{cases}$$

If P is a partition, then every subrectangle S will contain points (x,y) with x rational, and also points (x,y) with x irrational. Hence $m_S(f)=0$ and $M_S(f)=1$, so

$$L(f,P) = \sum_S 0\cdot v(S) = 0$$
$$U(f,P) = \sum_S 1\cdot v(S) = v([0,1]\times[0,1]) = 1.$$

Therefore f is not integrable.

23.2 Measure Zero and Content Zero

Definition 23.5 (Measure zero). We say $A\subset\mathbb{R}^n$ has ***measure*** 0, if for every $\varepsilon>0$ there exists a cover $\{U_1,U_2,\dots\}$ of A by closed rectangles such that

$$\sum_{n=1}^{\infty} v(U_n) < \varepsilon.$$

It is obvious (but nevertheless useful to remember) that if A has measure 0 and $B \subset A$, then B has measure 0. The reader may verify that open rectangles may be used instead of closed rectangles in the definition of measure 0.

Example. A set with only finitely many points clearly has measure 0.

Lemma 23.6. *If A has countably many points, then A also has measure 0.*

Lemma 23.7. *If $A = A_1 \cup A_2 \cup \cdots$ and each A_n has measure 0, then A has measure 0.*

Definition 23.8 (Content zero)**.**

23.3 Integrable Functions

23.4 Fubini's Theorem

23.5 Partitions of Unity

23.6 Change of Variables

VI
Complex Analysis

The starting point of our study is the idea of extending a function initially given for real values of the argument to one that is defined when the argument is complex. Thus, here the central objects are functions from the complex plane to itself

$$f\colon \mathbb{C} \to \mathbb{C},$$

or more generally, complex-valued functions defined on open subsets of $\mathbb{C}$.

24 Complex Functions

24.1 The Complex Plane

Basic Topology

$\mathbb{C}$ is a metric space, with metric $d(z,w) = |z-w|$. Hence all notions defined for general metric spaces, as outlined in Chapters 15 to 17 and 21, are applicable to $\mathbb{C}$.

If $z_0 \in \mathbb{C}$ and $r > 0$, we define the ***open disc*** of radius r centered at z_0 to be

$$D_r(z_0) = \{z \in \mathbb{C} \mid |z-z_0| < r\}.$$

The ***closed disc*** $D_r(z_0)$ of radius r centered at z_0 is defined by

$$D_r(z_0) = \{z \in \mathbb{C} \mid |z-z_0| \leq r\}.$$

The ***boundary*** of either the open or closed disc is the circle

$$C_r(z_0) = \{z \in \mathbb{C} \mid |z-z_0| = r\}.$$

Since the ***unit disc*** (the open disc centered at the origin and of radius 1) plays an important role, we will often denote it by $\mathbb{D}$:

$$\mathbb{D} = \{z \in \mathbb{C} \mid |z| < 1\}.$$

The last notion we need is that of connectedness. An open set $\Omega \subset \mathbb{C}$ is ***connected*** if it is not possible to find two disjoint non-empty open sets Ω_1 and Ω_2 such that

$$\Omega = \Omega_1 \cup \Omega_2.$$

A connected open set in $\mathbb{C}$ is called a ***region***. Similarly, a closed set F is connected if one cannot write $F = F_1 \cup F_2$ where F_1 and F_2 are disjoint non-empty closed sets.

There is an equivalent definition of connectedness for open sets in terms of curves, which is often useful in practice: an open set Ω is connected if and only if any two points in Ω can be

joined by a curve γ entirely contained in Ω.

Spherical Representation

For many purposes it is useful to extend the system $\mathbb{C}$ of complex numbers by introduction of a symbol ∞ to represent infinity. We call $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ the ***extended complex plane***.

We can define arithmetic operations on $\overline{\mathbb{C}}$.

- For $z \in \mathbb{C}$, define addition by
$$z + \infty = \infty.$$
Note that $\infty + \infty$ and $\infty - \infty$ are left undefined.
- For non-zero $z \in \mathbb{C}$, define multiplication by
$$z \times \infty = \infty,$$
with $\infty \times \infty = \infty$. The product $0 \times \infty$ is left undefined.
- Unlike the complex numbers, $\overline{\mathbb{C}}$ is not a field, since ∞ does not have an additive nor multiplicative inverse. Nonetheless, it is customary to define division on $\overline{\mathbb{C}}$ by
$$\frac{z}{0} = \infty, \quad \frac{z}{\infty} = 0$$
for all non-zero $z \in \mathbb{C}$, with $\infty/0 = \infty$ and $0/\infty = 0$. The quotients $0/0$ and ∞/∞ are left undefined.

One model of the extended complex plane is the *Riemann sphere*. Consider the unit sphere $\mathbb{S}$ defined by ${x_1}^2 + {x_2}^2 + {x_3}^2 = 1$. With every point on S, except $(0,0,1)$, we can associate a complex number
$$z = \frac{x_1 + ix_2}{1 - x_3}$$
and this correspondence is one to one. Indeed, from (24) we obtain
$$|z|^2 = \frac{x_1^2 + x_2^2}{(1-x_3)^2} = \frac{1+x_3}{1-x_3}$$
and hence
$$x_3 = \frac{|z|^2 - 1}{|z|^2 + 1}.$$
Further computation yields
$$x_1 = \frac{z + \bar{z}}{1 + |z|^2}, \quad x_2 = \frac{z - \bar{z}}{i(1 + |z|^2)}.$$

The correspondence can be completed by letting the point at infinity correspond to $(0,0,1)$, and we can thus regard the sphere as a representation of the extended plane. We note that the hemisphere $x_3 < 0$ corresponds to the disk $|z| < 1$ and the hemisphere $x_3 > 0$ to its outside $|z| > 1$.

24.2 Functions on The Complex Plane

We consider functions of the form $f\colon \mathbb{C} \to \mathbb{C}$. We can write $f = u + iv$ where $u, v\colon \mathbb{R}^2 \to \mathbb{R}$ are real-valued functions.

Limits

Recall the definition of a limit: we say $f\colon \Omega \subset \mathbb{C} \to \mathbb{C}$ has a limit w_0 at z_0 if

$$\forall \varepsilon > 0, \quad \exists \delta > 0, \quad \forall z \in \Omega, \quad 0 < |z - z_0| < \delta \implies |f(z) - w_0| < \varepsilon.$$

All the familiar limits results hold. Suppose $\lim_{z \to z_0} f(z) = w_1$ and $\lim_{z \to z_0} g(z) = w_2$.

(i) $\lim_{z \to z_0} f(z) \pm g(z) = w_1 \pm w_2$.

(ii) $\lim_{z \to z_0} f(z)g(z) = w_1 w_2$.

(iii) $\lim_{z \to z_0} \dfrac{f(z)}{g(z)} = \dfrac{w_1}{w_2}$ if $g(z), w_2 \neq 0$.

The complex conjugate is $\lim_{z \to z_0} \bar{z} = \bar{z_0}$.

For polynomials, $\lim_{z \to z_0} P(z) = P(z_0)$.

Method of proving $\lim_{z \to z_0} f(z) = w$:

1. Use limit rules and basic limit results.
2. Convert the complex function into two functions $\mathbb{R}^2 \to \mathbb{R}$.

To prove a limit does not exist, convert the complex function into two functions $\mathbb{R}^2 \to \mathbb{R}$ (or just apply directly on $\mathbb{C} \to \mathbb{C}$) and show that if we approach the target point from different directions we get different results.

Holomorphic Functions

We present the complex analogue of differentiability, which, at first glance, seems no different from the real case.

Definition 24.1. Let $\Omega \subset \mathbb{C}$ be open. We say $f\colon \Omega \to \mathbb{C}$ is ***holomorphic*** at $z_0 \in \Omega$ if

$$\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} \tag{24.1}$$

exists.

The value of the limit in (24.1) is known as the *derivative* of f at z_0; we write

$$f'(z_0) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}.$$

If $z = z_0 + h$ for some $h \in \mathbb{C}$, we can rewrite the above equation as

$$f'(z_0) = \lim_{h \to 0} \frac{f(z_0 + h) - f(z_0)}{h}.$$

It should be emphasised that h is a complex number that may approach 0 from any direction.

We can also rewrite

$$f(z_0 + h) - f(z_0) - f'(z_0)h = h\varepsilon(h), \tag{24.2}$$

where $\varepsilon(h) \to 0$ as $h \to 0$.

If f is holomorphic at every point of Ω, we say f is *holomorphic on* Ω. If $C \subset \mathbb{C}$ is closed, we say f is holomorphic on C if f is holomorphic in some open set containing C. Finally, we say f is ***entire*** if f is holomorphic on $\mathbb{C}$.

Lemma 24.2. *If f is holomorphic at z_0, then f is continuous at z_0.*

Proof. Suppose f is holomorphic at z_0. Then the limit $\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}$ exists. Thus

$$\begin{aligned} \lim_{z \to z_0} f(z) - f(z_0) &= \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}(z - z_0) \\ &= \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} \cdot \lim_{z \to z_0} (z - z_0) \\ &= f'(z_0) \cdot 0 = 0. \end{aligned}$$

Hence $\lim_{z \to z_0} f(z) = f(z_0)$, so f is continuous at z_0. □

The following lemma collects the basic facts about holomorphic functions. We omit the proof, which is essentially identical to the real case.

Lemma 24.3. *Suppose $\Omega \subset \mathbb{C}$ is open, and $f, g \colon \Omega \to \mathbb{C}$ are holomorphic on Ω.*

(i) $f + g$ is holomorphic on Ω, and *(sums)*

$$(f + g)' = f' + g'.$$

(ii) fg is holomorphic on Ω, and *(products)*

$$(fg)' = f'g + fg'.$$

(iii) f/g is holomorphic on Ω (provided $g(z) \neq 0$), and *(quotients)*

$$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}.$$

Lemma 24.4 (Chain rule). *Suppose Ω and U are open subsets of $\mathbb{C}$, and $f\colon \Omega \to U$ and $g\colon U \to \mathbb{C}$ are holomorphic. Then*

$$(g \circ f)'(z) = g'(f(z))f'(z) \quad (z \in \Omega). \tag{24.3}$$

Example (Polynomials). $f(z) = z$ is holomorphic on any open set in $\mathbb{C}$, and $f'(z) = 1$. In fact, any polynomial

$$p(z) = a_n z^n + \cdots + a_1 z + a_0$$

is holomorphic in the entire complex plane, and

$$p'(z) = n a_n z^{n-1} + \cdots + a_1.$$

Example. $f(z) = \frac{1}{z}$ is holomorphic on any open set in $\mathbb{C}$ that does not contain the origin, and $f'(z) = -\frac{1}{z^2}$.

Example. $f(z) = \bar{z}$ is not holomorphic. Indeed, we have

$$\frac{f(z_0 + h) - f(z_0)}{h} = \frac{\bar{h}}{h}$$

which has no limit as $h \to 0$, as one can see by first taking h real and then h purely imaginary.

Cauchy–Riemann Equations

To each complex-valued function $f = u + iv$, we associate the mapping $F(x,y) = (u(x,y), v(x,y))$ from $\mathbb{R}^2$ to $\mathbb{R}^2$.

Recall from Chapter 22 that a function $F(x,y) = (u(x,y), v(x,y))$ is said to be *differentiable* at a

point $P_0 = (x_0, y_0)$ if there exists a linear transformation $J\colon \mathbb{R}^2 \to \mathbb{R}^2$ such that

$$\lim_{|h|\to 0} \frac{F(P_0+h) - F(P_0) - J(h)}{|h|} = 0.$$

Equivalently, we can write

$$F(P_0+h) - F(P_0) = J(h) + |h|\varepsilon(h),$$

with $\|\varepsilon(h)\| \to 0$ as $|h| \to 0$. The linear transformation J is unique and is called the *derivative* of F at P_0.

If F is differentiable, the partial derivatives of u and v exist, and the linear transformation J can be described in the standard basis of $\mathbb{R}^2$ by the Jacobian matrix of F:

$$J = J_F(x,y) = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix}.$$

In the case of complex differentiation, the derivative is a complex number $f'(z_0)$; in the case of real derivatives, it is a matrix. There is, however, a connection between these two notions, which is given in terms of special relations that are satisfied by the entries of the Jacobian matrix, that is, the partials of u and v.

Theorem 24.5 (Cauchy–Riemann equations)**.** *Suppose f is holomorphic. Then its real and imaginary parts satisfy*

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}. \tag{24.4}$$

Proof. Consider the limit

$$f'(z_0) = \lim_{h\to 0} \frac{f(z_0+h) - f(z_0)}{h}$$

when h is first real, say $h = h_1 + ih_2$ with $h_2 = 0$. Then, if we write $z = x+iy$, $z_0 = x_0 + iy_0$, and $f(z) = f(x,y)$, we find that

$$\begin{aligned} f'(z_0) &= \lim_{h_1\to 0} \frac{f(x_0+h_1, y_0) - f(x_0, y_0)}{h_1} \\ &= \frac{\partial f}{\partial x}(z_0). \end{aligned}$$

Now taking h purely imaginary, say $h = ih_2$, a similar argument yields

$$\begin{aligned} f'(z_0) &= \lim_{h_2 \to 0} \frac{f(x_0, y_0 + h_2) - f(x_0, y_0)}{ih_2} \\ &= \frac{1}{i}\frac{\partial f}{\partial y}(z_0) = -i\frac{\partial f}{\partial y}(z_0). \end{aligned}$$

Hence

$$\frac{\partial f}{\partial x} = -i\frac{\partial f}{\partial y}.$$

Writing $f = u + iv$, we find after separating real and imaginary parts, that the partials of u and v exist, and they satisfy (24.4). □

Remark. This suggests that complex differentiability is a much more rigid property than one might think at first sight; if f is differentiable then these partial derivatives do exist, and moreover they are subject to a constraint.

We can clarify the situation further by defining two differential operators

$$\begin{aligned} \frac{\partial}{\partial z} &= \frac{1}{2}\left(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right), \\ \frac{\partial}{\partial \bar{z}} &= \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right). \end{aligned}$$

Proposition 24.6. *If f is holomorphic at z_0, then*

$$\frac{\partial f}{\partial \bar{z}} = 0$$

and thus

$$f'(z_0) = \frac{\partial f}{\partial z}(z_0) = 2\frac{\partial u}{\partial z}(z_0).$$

If we write $F(x,y) = f(z)$, then F is differentiable in the sense of real variables, and

$$\det J_F(x_0, y_0) = |f'(z_0)|^2.$$

Proof. Taking real and imaginary parts, it is easy to see that the Cauchy–Riemann equations are equivalent to $\frac{\partial f}{\partial \bar{z}} = 0$.

Moreover, by our earlier observation,

$$f'(z_0) = \frac{1}{2}\left(\frac{\partial f}{\partial x}(z_0) - i\frac{\partial f}{\partial y}(z_0)\right) = \frac{\partial f}{\partial z}(z_0),$$

and the Cauchy–Riemann equations give

$$\frac{\partial f}{\partial z} = 2\frac{\partial u}{\partial z}.$$

To prove that F is differentiable it suffices to observe that if $h = (h_1, h_2)$ and $h = h_1 + ih_2$, then the Cauchy–Riemann equations imply

$$J_F(x_0, y_0)(h) = \left(\frac{\partial u}{\partial x} - i\frac{\partial u}{\partial y}\right)(h_1 + ih_2) = f'(z_0)h,$$

where we have identified a complex number with the pair of real and imaginary parts. After a final application of the Cauchy–Riemann equations, the above results imply that

$$\det J_F(x_0, y_0) = \frac{\partial u}{\partial x}\frac{\partial v}{\partial y} - \frac{\partial v}{\partial x}\frac{\partial u}{\partial y} = \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 = \left|2\frac{\partial u}{\partial z}\right|^2 = |f'(z_0)|^2.$$

□

So far, we have assumed that f is holomorphic and deduced relations satisfied by its real and imaginary parts. The next result contains an important converse, which completes the circle of ideas presented here.

Proposition 24.7. *Suppose $f = u + iv$ is a complex-valued function defined on an open set $\Omega \subset \mathbb{C}$. If $u, v\colon \mathbb{R}^2 \to \mathbb{R}$ are continuously differentiable and satisfy the Cauchy–Riemann equations on Ω, then f is holomorphic on Ω and $f'(z) = \frac{\partial f}{\partial z}$.*

Proof. Write

$$u(x + h_1, y + h_2) - u(x, y) = \frac{\partial u}{\partial x}h_1 + \frac{\partial u}{\partial y}h_2 + |h|\varepsilon_1(h)$$

and

$$v(x + h_1, y + h_2) - v(x, y) = \frac{\partial v}{\partial x}h_1 + \frac{\partial v}{\partial y}h_2 + |h|\varepsilon_2(h)$$

where the remainders $\varepsilon_1(h), \varepsilon_2(h) \to 0$ as $|h| \to 0$, and $h = h_1 + ih_2$. Using the Cauchy–Riemann equations we find that

$$f(z + h) - f(z) = \left(\frac{\partial u}{\partial x} - i\frac{\partial u}{\partial y}\right)(h_1 + ih_2) + |h|\varepsilon(h),$$

where $\varepsilon(h) = \varepsilon_1(h) + \varepsilon_2(h) \to 0$ as $|h| \to 0$. Hence f is holomorphic and

$$f'(z) = 2\frac{\partial u}{\partial z} = \frac{\partial f}{\partial z}.$$

□

Example. The function $f(z) = \bar{z}$ is not holomorphic.

Proof. Let $u, v : \mathbb{R}^2 \to \mathbb{R}$ be the components of f. Then $u(x,y) = x$, $v(x,y) = -y$, and so

$$\frac{\partial u}{\partial x} = 1, \quad \frac{\partial u}{\partial y} = 0, \quad \frac{\partial v}{\partial x} = 0, \quad \frac{\partial v}{\partial y} = -1.$$

Since u and v do not satisfy the Cauchy–Riemann equations, f is not holomorphic. □

We shall prove later that the derivative of an holomorphic function is itself holomorphic. By this fact, u and v have continuous second partial derivatives. Differentiating (24.4) again gives

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 v}{\partial x \partial y}, \quad \frac{\partial^2 u}{\partial y^2} = -\frac{\partial^2 v}{\partial y \partial x}.$$

Hence

$$\begin{aligned} \Delta u &= \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \\ \Delta v &= \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0. \end{aligned} \tag{24.5}$$

(24.5) is called *Laplace's equation*. Solutions to (24.5) are said to be ***harmonic***. If two harmonic functions u and v satisfy the Cauchy–Riemann equations (24.4), then v is called the *conjugate harmonic function* of u.

Power Series

Recall our discussion of power series in Chapter 21, including Hadamard's formula (21.1) for the radius of convergence of a power series.

Having defined complex differentiation, we now prove the complex analogue of 21.6, concerning the derivative of complex power series.

Proposition 24.8. *The power series $f(z) = \sum_{n=0}^{\infty} a_n z^n$ defines a holomorphic function in its disc of convergence. Then*

$$f'(z) = \sum_{n=0}^{\infty} n a_n z^{n-1}. \tag{24.6}$$

That is, the derivative of a power series is obtained by differentiating the series term-by-term.

Proof. The desired series $\sum_{n=1}^{\infty} n a_n z^{n-1}$ has the same radius of convergence, because $\sqrt[n]{n} \to 1$. Proof: Set $\sqrt[n]{n} = 1 + \delta_n$. Then $\delta_n > 0$, and by use of the binomial theorem $(1+\delta_n)^n > 1 + \frac{1}{2}n(n-1)\delta_n^2$. This gives $\delta_n^2 < 2/n$, and hence $\delta_n \to 0$.

For $|z| < R$ we shall write

$$f(z) = \sum_{n=0}^{\infty} a_n z^n = s_n(z) + R_n(z)$$

where

$$s_n(z) = \sum_{k=0}^{n-1} a_k z^k, \quad R_n(z) = \sum_{k=n}^{\infty} a_k z^k$$

and also

$$f_1(z) = \sum_{n=1}^{\infty} n a_n z^{n-1} = \lim_{n\to\infty} s_n'(z).$$

We have to show that $f'(z) = f_1(z)$.

Consider the identity

$$\frac{f(z)-f(z_0)}{z-z_0} - f_1(z_0) = \left(\frac{s_n(z)-s_n(z_0)}{z-z_0} - s_n'(z_0)\right) + (s_n'(z_0) - f_1(z_0)) + \left(\frac{R_n(z_0)-R_n(z_0)}{z-z_0}\right)$$

where we assume $z \neq z_0$. The last term can be rewritten as

$$\sum_{k=n}^{\infty} a_k (z^{k-1} + z^{k-2}z_0 + \cdots + z z_0^{k-2} + z_0^{k-1}),$$

and we conclude that

$$\left|\frac{R_n(z_0)-R_n(z_0)}{z-z_0}\right| \leq \sum_{k=n}^{\infty} k|a_k|R^{k-1}.$$

The expression on the right is the remainder term in a convergent series. Hence we can find $N_1 \in \mathbb{N}$ such that

$$\left|\frac{R_n(z_0)-R_n(z_0)}{z-z_0}\right| < \frac{\varepsilon}{3}$$

for all $n \geq N_1$.

There also exists $N_1 \in \mathbb{N}$ such that

$$\left|s_n'(z_0) - f_1(z_0)\right| < \frac{\varepsilon}{3}$$

for all $n \geq N_2$.

Choose a fixed $n \geq N_1, N_2$. By the definition of derivative we can find $\delta > 0$ such that $0 < |z-z_0| < \delta$ implies

$$\left|\frac{s_n(z)-s_n(z_0)}{z-z_0} - s_n'(z_0)\right| < \frac{\varepsilon}{3}.$$

When all these inequalities are combined it follows that

$$\left|\frac{f(z)-f(z_0)}{z-z_0}\right| < \varepsilon$$

when $0 < |z-z_0| < \delta$. Hence we have proved that $f'(z_0)$ exists and equals $f_1(z_0)$. □

Repeated differentiation of $f(z)$ shows that a power series is infinitely complex differentiable in

its disc of convergence, and its derivatives are given explicitly by

$$f^{(k)}(z) = \sum_{n=k}^{\infty} n(n-1)\cdots(n-k+1)a_n z^{n-k}.$$

In particular, we see that $a_k = \frac{f^{(k)}(0)}{k!}$, and the power series becomes

$$f(z) = f(0) + f'(0)z + \frac{f''(0)}{2!}z^2 + \cdots + \frac{f^{(n)}(0)}{n!}z^n + \cdots \tag{24.7}$$

This is the familiar *Taylor–Maclaurin development*.

We have proved it only under the assumption that $f(z)$ has a power series development. We do know that the development is uniquely determined, if it exists, but the main part is still missing, namely that every analytic function has a Taylor development.

Definition 24.9 (Analytic function)**.** Let $\Omega \subset \mathbb{C}$. We say $f\colon \Omega \to \mathbb{C}$ is ***analytic*** at $z_0 \in \Omega$ if there exist $c_0, c_1, \cdots \in \mathbb{C}$ such that

$$f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n$$

and the series is convergent to $f(z)$ for all z in a neighbourhood of z_0.

If f is analytic at every point in Ω, we say f is *analytic on* Ω.

The Exponential and Trigonometric Functions

Recall the exponential function as defined in (21.6).

We are familiar with the exponential function e^x of a real variable, which has the property that $(e^x)' = e^x$. The complex exponential has the same property:

Lemma 24.10. $\exp'(z) = \exp(z)$ *for all* $z \in \mathbb{C}$.

Proof. Using 24.8, we calculate the derivative of $\exp(z)$ by differentiating term-by-term:

$$\exp'(z) = (1)' + \sum_{n=1}^{\infty} \left(\frac{z^n}{n!}\right)' = \sum_{n=1}^{\infty} \frac{nz^{n-1}}{n!} = \sum_{n=1}^{\infty} \frac{z^{n-1}}{(n-1)!} = \exp(z).$$

□

Recall the trigonometric functions sine and cosine as defined in (21.10). Having defined complex differentiation, it follows that

$$\cos' z = -\sin z, \quad \sin' z = \cos z.$$

24.3 Holomorphic Functions as Mappings

Conformality

Definition 24.11 (Curve). A (parametrised) curve is a continuous map $\gamma\colon [a,b] \to \mathbb{C}$. We also say

- γ is *simple* if γ is injective on $[a,b]$ (not self-intersecting);
- γ is *closed* if $\gamma(a) = \gamma(b)$ (end points coincide);
- γ is a *simple closed* curve (or *Jordan curve*) if γ is closed and $\gamma|_{[a,b)}$ is simple.

We say a curve γ is ***smooth*** if $\gamma'(t)$ exists and is continuous on $[a,b]$, and $\gamma'(t) \neq 0$ for $t \in [a,b]$.

Remark. At the endpoints $t = a$ and $t = b$, the quantities $z'(a)$ and $z'(b)$ are interpreted as the one-sided limits

$$z'(a) = \lim_{h\to 0^+} \frac{z(a+h)-z(a)}{h}, \quad z'(b) = \lim_{h\to 0^-} \frac{z(b+h)-z(b)}{h},$$

which are the right-hand derivative of $z(t)$ at a, and the left-hand derivative of $z(t)$ at b, respectively.

Definition 24.12. A curve $\gamma\colon [a,b] \to \mathbb{C}$ is ***piecewise-smooth*** if γ is continuous on $[a,b]$ and there exist points

$$a = t_0 \leq t_1 \leq \cdots \leq t_n = b,$$

where $\gamma(t)$ is smooth in the intervals $[t_i, t_{i+1}]$.

For brevity, we shall call any piecewise-smooth curve a *curve*, since these will be the objects we shall be primarily concerned with.

We say two parametrisations

$$z\colon [a,b] \to \mathbb{C}, \quad \overline{z}\colon [c,d] \to \mathbb{C}$$

are *equivalent* if there exists a continuously differentiable bijection $s \mapsto t(s)$ from $[c,d]$ to $[a,b]$ so that $t'(s) > 0$ and

$$\overline{z}(s) = z(t(s)).$$

Remark. The condition $t'(s) > 0$ says precisely that the orientation is preserved: as s travels from c to d, then $t(s)$ travels from a to b.

The family of all parametrisations that are equivalent to $z(t)$ determines a smooth curve $\gamma \subset \mathbb{C}$, namely the image of $[a,b]$ under z with the orientation given by z as t travels from a to b.

Since γ carries an orientation, it is natural to say that γ begins at $z(a)$ and ends at $z(b)$.

A basic example consists of a circle. Consider the circle $C_r(z_0)$ centered at z_0 and of radius r:

$$C_r(z_0) = \{z \in \mathbb{C} \mid |z - z_0| = r\}.$$

Definition 24.13 (Orientation)**.** The positive orientation (counterclockwise) is the one that is given by the standard parametrisation

$$z(t) = z_0 + re^{it} \quad (0 \leq t \leq 2\pi).$$

The negative orientation (clockwise) is given by

$$z(t) = z_0 + re^{-it} \quad (0 \leq t \leq 2\pi).$$

In the following chapters, we shall denote by C a general positively oriented circle.

2.1 Arcs and Closed Curves 2.2 Analytic Functions in Regions 2.3 Conformal Mapping 2.4 Length and Area

Linear Transformations

3.1 The Linear Group 3.2 The Cross Ratio 3.3 Symmetry 3.4 Oriented Circles 3.5 Families of Circles

Elementary Conformal Mappings

4.1 The Use of Level Curves 4.2 A Survey of Elementary Mappings 4.3 Elementary Riemann Surfaces

25 Complex Integration

25.1 Fundamental Theorems

Line Integrals

The most immediate generalisation of a real integral is to the definite integral of a complex function over a real interval. If $f(t) = u(t) + iv(t)$ is a continuous function defined on (a,b), we define

$$\int_a^b f(t)\,\mathrm{d}t = \int_a^b u(t)\,\mathrm{d}t + i\int_a^b v(t)\,\mathrm{d}t\,.$$

This integral has most of the properties of the real integral.

Definition 25.1 (Integral along curve). Given a piecewise differentiable curve γ in $\mathbb{C}$ parametrised by $\gamma\colon [a,b] \to \mathbb{C}$, and a function defined on γ, define the ***integral*** of f along γ by

$$\int_\gamma f(z)\,\mathrm{d}z := \int_a^b f\left(z(t)\right) z'(t)\,\mathrm{d}t\,. \tag{25.1}$$

For the integral to be well-defined, we must show that the integral on the RHS is independent of the parametrisation chosen for γ. Suppose $\bar{z}$ is an equivalent parametrisation as above. Then the change of variables formula and the chain rule imply that

$$\int_a^b f\left(z(t)\right) z'(t)\,\mathrm{d}t = \int_c^d f\left(z(t(s))\right) z'\left(t(s)\right) t'(s)\,\mathrm{d}s = \int_c^d f\left(\bar{z}(s)\right) \bar{z}'(s)\,\mathrm{d}s\,.$$

This proves that the integral of f over γ is well defined.

If γ is piecewise smooth, then the integral of f over γ is simply the sum of the integrals of f over the smooth parts of γ, so if $z(t)$ is a piecewise-smooth parametrisation as before, then

$$\int_\gamma f(z)\,\mathrm{d}z = \sum_{i=1}^{n} \int_{a_{i-1}}^{a_i} f\left(z(t)\right) z'(t)\,\mathrm{d}t\,.$$

Lemma 25.2 (Basic properties).

(i) Linearity: if $\alpha, \beta \in \mathbb{C}$, then

$$\int_\gamma (\alpha f(z) + \beta g(z))\,dz = \alpha \int_\gamma f(z)\,dz + \beta \int_\gamma g(z)\,dz.$$

(ii) If γ^- is γ with the reverse orientation, then

$$\int_\gamma f(z)\,dz = -\int_{\gamma^-} f(z)\,dz.$$

(iii) One has the inequality

$$\left|\int_\gamma f(z)\,dz\right| \leq \sup_{z\in\gamma} |f(z)| \cdot \Lambda(\gamma).$$

A ***primitive*** for f on Ω is a function F that is holomorphic on Ω and such that $F'(z) = f(z)$ for all $z \in \Omega$.

Proposition 25.3. *If a continuous function f has a primitive F in Ω, and γ is a curve in Ω that begins at w_1 and ends at w_2, then*

$$\int_\gamma f(z)\,dz = F(w_2) - F(w_1). \tag{25.2}$$

Corollary 25.4. *If γ is a closed curve in an open set Ω, and f is continuous and has a primitive in Ω, then*

$$\int_\gamma f(z)\,dz = 0.$$

Proof. This is immediate since the end-points of a closed curve coincide. □

Corollary 25.5. *If f is holomorphic in a region Ω and $f' = 0$, then f is constant.*

Rectifiable Arcs

By definition, the ***length*** of the smooth curve γ is

$$\Lambda(\gamma) = \int_a^b |z'(t)|\,dt.$$

Arguing as we just did, it is clear that this definition is also independent of the parametrisation. Also, if γ is only piecewise-smooth, then its length is the sum of the lengths of its smooth parts.

Line Integrals as Functions of Arcs

Cauchy's Theorem for a Rectangle

Cauchy's Theorem in a Disk

25.2 Cauchy's Integral Formula

2.1 The Index of a Point with Respect to a Closed Curve 2.2 The Integral Formula 2.3 Higher Derivatives

25.3 Local Properties of Analytical Functions

3.1 Removable Singularities. Taylor's Theorem 3.2 Zeros and Poles 3.3 The Local Mapping 3.4 The Maximum Principle

25.4 The General Form of Cauchy's Theorem

4.1 Chains and Cycles 4.2 Simple Connectivity 4.3 Homology 4.4 The General Statement of Cauchy's Theorem 4.5 Proof of Cauchy's Theorem 4.6 Locally Exact Differentials 4.7 Multiply Connected Regions

25.5 The Calculus of Residues

5.1 The Residue Theorem 5.2 The Argument Principle 5.3 Evaluation of Definite Integrals

VII
General Topology

The study of topology simultaneously simplifies and generalises the theory of metric spaces. By discarding the metric, and focusing solely on the more basic and fundamental notion of an open set, many arguments and proofs are simplified. And many constructions (such as the important concept of a quotient space) cannot be carried out in the setting of metric spaces: they need the more general framework of topological spaces.

26 Topological Spaces and Continuous Functions

26.1 Topologies

Definitions and Examples

Definition 26.1 (Topological space)**.** Let X be a set. We say $\mathscr{T} \subset \mathscr{P}(X)$ is a ***topology*** on X if

(i) $\emptyset, X \in \mathscr{T}$;

(ii) if $\{U_i\}_{i\in I}$ are in $\mathscr{T}$, then $\bigcup_{i\in I} U_i \in \mathscr{T}$; (closed under arbitrary unions)

(iii) if $U_1, \ldots, U_n \in \mathscr{T}$, then $\bigcap_{i=1}^{n} U_i \in \mathscr{T}$. (closed under finite intersections)

If $\mathscr{T}$ is a topology on X, we say $(X, \mathscr{T})$ is a ***topological space***. The elements of $\mathscr{T}$ are called *open sets* of X.

Notation. If the topology $\mathscr{T}$ is clear, we simply omit it and denote a topological space as X.

Example. Let X be any non-empty set.

- The *discrete topology* on X is the set of all subsets of X; that is, $\mathscr{T} = \mathscr{P}(X)$. (Every subset of X is open.)
- The *trivial topology* on X is $\mathscr{T} = \{\emptyset, X\}$.
- The *co-finite topology* on X consists of the empty set together with every subset U of X such that U^c is finite:

$$U \in \mathscr{T} \iff U = \emptyset \quad \text{or} \quad U^c \text{ is finite.}$$

- Any metric space (X, d), with $\mathscr{T}$ equal to the collection of all subsets of X that

are open in the metric space sense. This topology is called the *metric topology* on X.

Definition 26.2. Suppose $\mathscr{T}$ and $\mathscr{T}'$ are two topologies on a given set X. We say that

(i) $\mathscr{T}$ is ***finer*** than $\mathscr{T}'$ if $\mathscr{T} \supset \mathscr{T}'$;

(ii) $\mathscr{T}$ is ***coarser*** than $\mathscr{T}'$ if $\mathscr{T} \subset \mathscr{T}'$;

(iii) $\mathscr{T}$ is ***comparable*** with $\mathscr{T}'$ if either $\mathscr{T} \supset \mathscr{T}'$ or $\mathscr{T} \subset \mathscr{T}'$.

Example. The indiscrete topology is the coarsest topology possible, while the discrete topology is the finest topology possible.

Bases

In linear algebra, every vector space is generated by a basis. In topology, we have a similar notion, as it is usually hard to define a topology by specifying all the open sets.

Definition 26.3 (Basis)**.** We say $\mathscr{B} \subset \mathscr{P}(X)$ is a ***basis*** for a topology on X if

(i) for all $x \in X$, there exists $B \in \mathscr{B}$ such that $x \in B$;

(ii) for all $B_1, B_2 \in \mathscr{B}$ and $x \in B_1 \cap B_2$, there exists $B_3 \in \mathscr{B}$ such that $x \in B_3 \subset B_1 \cap B_2$.

The members of a basis are called *basis elements*.

Property (i) states that the elements of $\mathscr{B}$ cover X. Property (ii) can be visualised as follows:

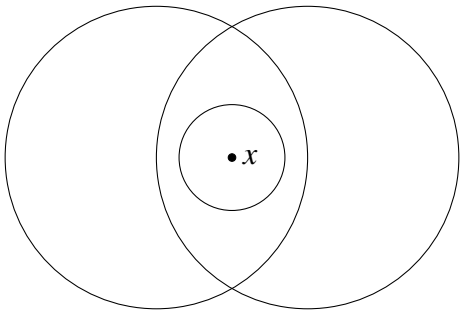


Figure 26.1: Property (ii) of Definition 26.3

We can use a basis to *generate* a topology.

Definition 26.4 (Topology generated by basis)**.** We define the *topology $\mathscr{T}$ generated by basis $\mathscr{B}$* as follows:

$$U \in \mathscr{T} \iff \forall x \in U, \quad \exists B \in \mathscr{B}, \quad x \in B \subset U.$$

Lemma. *The collection $\mathscr{T}$ generated by the basis $\mathscr{B}$ is a topology on X.*

Proof.

(i) $\emptyset$ satisfies the defining condition of openness vacuously, so $\emptyset \in \mathscr{T}$.

$X \in \mathscr{T}$ follows from (i) of Definition 26.3.

(ii) Let $\{U_i\}_{i \in I}$ be a collection of elements of $\mathscr{T}$. We want to show $U = \bigcup_{i \in I} U_i \in \mathscr{T}$.

Let $x \in U$. Then $x \in U_i$ for some $i \in I$. Since $U_i \in \mathscr{T}$, there exists $B \in \mathscr{B}$ such that $x \in B \subset U_i$. Thus $x \in B \subset U$, so $U \in \mathscr{T}$.

(iii) Let $U_1, U_2 \in \mathscr{T}$. We want to show $U_1 \cap U_2 \in \mathscr{T}$.

Let $x \in U_1 \cap U_2$. Since $U_1 \in \mathscr{T}$, there exists $B_1 \in \mathscr{B}$ such that $x \in B_1 \subset U_1$; since $U_2 \in \mathscr{T}$, there exists $B_2 \in \mathscr{B}$ such that $x \in B_2 \subset U_2$. Then $x \in B_1 \cap B_2$.

Since $\mathscr{B}$ is a basis, by (ii) of Definition 26.3, there exists $B_3 \in \mathscr{B}$ such that $x \in B_3 \subset B_1 \cap B_2$. Thus $U_1 \cap U_2 \in \mathscr{T}$.

Finally, we show by induction that any finite intersection $U_1 \cap \cdots \cap U_n \in \mathscr{T}$. This is trivial for $n = 1$; suppose it true for $n-1$ and prove it for n. Now

$$(U_1 \cap \cdots \cap U_n) = (U_1 \cap \cdots \cap U_{n-1}) \cap U_n.$$

By hypothesis, $U_1 \cap \cdots \cap U_{n-1} \in \mathscr{T}$. Thus the intersection of $U_1 \cap \cdots \cap U_{n-1}$ and U_n also belongs to $\mathscr{T}$.

□

Another way of describing the topology generated by a basis is given in the following result:

Lemma 26.5. *Let $\mathscr{T}$ be the topology on X generated by basis $\mathscr{B}$. Then $\mathscr{T}$ equals the collection of all unions of elements of $\mathscr{B}$.*

Proof. Let $\mathscr{B} = \{B_i \mid i \in I\}$.

$\boxed{\supset}$ If $B_i \in \mathscr{B}$, see that

$$\forall x \in B_i, \quad x \in B_i \subset B_i \implies B_i \in \mathscr{T}.$$

Since $\mathscr{T}$ is a topology, the arbitrary unions of B_i's must be in $\mathscr{T}$.

$\boxed{\subset}$ Let $U \in \mathscr{T}$. Then for each $x \in U$, there exists $B_x \in \mathscr{B}$ such that $x \in B_x \subset U$. Then $U = \bigcup_{x \in U} B_x$, so U is a union of elements of $\mathscr{B}$. □

Remark. The above result states that every $U \in \mathscr{T}$ can be expressed as a union of basis elements.

Given a basis, we can construct a topology. Now we go in the reverse direction: given a topology, construct a basis.

Lemma 26.6. *Let $(X, \mathscr{T})$ be a topological space. Suppose $\mathscr{C} \subset \mathscr{T}$ is such that*

$$\forall U \in \mathscr{T}, \quad \forall x \in U, \quad \exists C \in \mathscr{C}, \quad x \in C \subset U.$$

Then $\mathscr{C}$ is a basis for $\mathscr{T}$.

Proof. We first show that $\mathscr{C}$ is a basis.

(i) For all $x \in X$, since $X \in \mathscr{T}$, by hypothesis, there exists $C \in \mathscr{C}$ such that $x \in C \subset X$.

(ii) Let $x \in C_1 \cap C_2$, where $C_1, C_2 \in \mathscr{C} \subset \mathscr{T}$. Thus $C_1, C_2 \in \mathscr{T}$, so $C_1 \cap C_2 \in \mathscr{T}$. By hypothesis, there exists $C_3 \in \mathscr{C}$ such that $x \in C_3 \subset C_1 \cap C_2$.

Let $\mathscr{T}'$ be the topology generated by $\mathscr{C}$; that is,

$$U \in \mathscr{T}' \iff \forall x \in U, \quad \exists C \in \mathscr{C}, \quad x \in C \subset U.$$

We will show that $\mathscr{T} = \mathscr{T}'$.

$\boxed{\subset}$ Let $U \in \mathscr{T}$, $x \in U$. By hypothesis, there exists $C \in \mathscr{C}$ such that $x \in C \subset U$. By definition, $U \in \mathscr{T}'$. Hence $\mathscr{T} \subset \mathscr{T}'$.

$\boxed{\supset}$ Conversely, let $W \in \mathscr{T}'$. By 26.5, W is a union of elements of $\mathscr{C}$. Since each element of $\mathscr{C}$ is an element of $\mathscr{T}$ (and thus open), and a union of open sets is open, we have $W \in \mathscr{T}$. Hence $\mathscr{T}' \subset \mathscr{T}$. □

When topologies are given by bases, the next result is a criterion to determine whether one topology is finer than another.

Lemma 26.7. *Let $\mathscr{B}$ and $\mathscr{B}'$ be bases for the topologies $\mathscr{T}$ and $\mathscr{T}'$ respectively on X. Then the following are equivalent:*

(i) $\mathscr{T}'$ is finer than $\mathscr{T}$.

(ii) For all $x \in X$, and for all $B \in \mathscr{B}$ such that $x \in B$, there exists $B' \in \mathscr{B}'$ such that $x \in B' \subset B$.

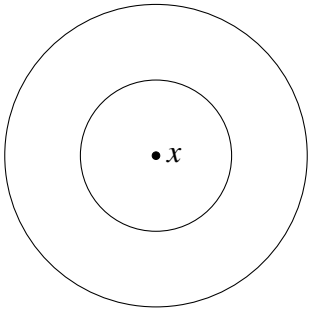


Figure 26.2: (ii) of 26.7

Proof.

$\boxed{\text{(ii)} \Longrightarrow \text{(i)}}$ Let $U \in \mathscr{T}$. To show that $\mathscr{T} \subset \mathscr{T}'$, we want to show that $U \in \mathscr{T}'$.

Let $x \in U$. Since $\mathscr{B}$ generates $\mathscr{T}$, there exists $B \in \mathscr{B}$ such that $x \in B \subset U$. By (ii), there exists $B' \in \mathscr{B}'$ such that $x \in B' \subset B$. Then $x \in B' \subset U$, so $U \in \mathscr{T}'$, by definition.

$\boxed{\text{(i)} \Longrightarrow \text{(ii)}}$ We are given $x \in X$ and $B \in \mathscr{B}$, with $x \in B$.

Now $B \in \mathscr{T}$ by definition, and $\mathscr{T} \subset \mathscr{T}'$ by (i); therefore, $B \in \mathscr{T}'$. Since $\mathscr{T}'$ is generated by $\mathscr{B}'$, there exists $B' \in \mathscr{B}'$ such that $x \in B' \subset B$. □

We now define three topologies on the real line $\mathbb{R}$, all of which are of interest.

Definition 26.8.

(i) Let $\mathscr{B}$ be the collection of all open intervals in $\mathbb{R}$. The topology generated by $\mathscr{B}$ is called the ***standard topology*** on $\mathbb{R}$.

Whenever we consider $\mathbb{R}$, we shall suppose it is given this topology unless stated otherwise.

(ii) Let $\mathscr{B}'$ be the collection of all half-open intervals of the form $[a,b)$. The topology generated by $\mathscr{B}'$ is called the ***lower limit topology*** on $\mathbb{R}$.

When $\mathbb{R}$ is given the lower limit topology, we denote it by $\mathbb{R}_\ell$.

(iii) Let $K = \{\frac{1}{n} \mid n \in \mathbb{Z}^+\}$, and let $\mathscr{B}''$ be the collection of all open intervals (a,b), along with all sets of the form $K \setminus (a,b)$. The topology generated by B'' is called the ***K-topology*** on $\mathbb{R}$.

When $\mathbb{R}$ is given this topology, we denote it by $\mathbb{R}_K$.

It is easy to see that all three of these collections are bases; in each case, the intersection of two basis elements is either another basis element or is empty. The relation between these topologies is the following:

Lemma 26.9. *The topologies of $\mathbb{R}_\ell$ and $\mathbb{R}_K$ are strictly finer than the standard topology on $\mathbb{R}$, but are not comparable with one another.*

Proof. Let $\mathscr{T}$, $\mathscr{T}'$, and $\mathscr{T}''$ be the topologies of $\mathbb{R}$, $\mathbb{R}_\ell$, and $\mathbb{R}_K$, respectively.

Given a basis element (a,b) for $\mathscr{T}$ and a point x of (a,b), the basis element $[x,b)$ for $\mathscr{T}'$ contains x and lies in (a,b). On the other hand, given the basis element $[x,d)$ for $\mathscr{T}'$, there is no open interval (a,b) that contains x and lies in $[x,d)$. Thus $\mathscr{T} \subset \mathscr{T}'$, so $\mathscr{T}'$ is strictly finer than $\mathscr{T}$.

A similar argument applies to $\mathbb{R}_K$. Given a basis element (a,b) for $\mathscr{T}$ and a point x of (a,b), this same interval is a basis element for $\mathscr{T}''$ that contains x. On the other hand, given the basis element $B = (-1,1) \setminus K$ for $\mathscr{T}''$ and the point 0 of B, there is no open interval that contains 0 and lies in $\mathscr{B}$.

We leave it to you to show that the topologies of $\mathbb{R}_\ell$ and $\mathbb{R}_K$ are not comparable. □

Since the topology generated by a basis $\mathscr{B}$ may be described as the collection of arbitrary unions of elements of $\mathscr{B}$ (by 26.5), what happens if we start with a given collection of sets and take finite intersections of them as well as arbitrary unions? This leads to the notion of a *subbasis* for a topology.

Definition 26.10 (Subbasis)**.** A ***subbasis*** $\mathscr{S}$ for a topology on X is a collection of subsets of X whose union equals X.

Definition 26.11 (Topology generated by subbasis)**.** The *topology $\mathscr{T}$ generated by the subbasis $\mathscr{S}$* is defined as the collection of all unions of finite intersections of elements of $\mathscr{S}$:

$$U \in \mathscr{T} \iff U = \text{union of finite intersections in } \mathscr{S}.$$

Lemma. *The collection $\mathscr{T}$ generated by the subbasis $\mathscr{S}$ is a topology.*

Proof. Consider the collection

$$\mathscr{B} = \{\text{all finite intersections of elements of } \mathscr{S}\}.$$

It suffices to show that $\mathscr{B}$ is a basis, for then by 26.5, the collection $\mathscr{T}$ of all unions of elements of $\mathscr{B}$ is a topology.

(i) Let $x \in X$. Then x belongs to an element of $\mathscr{S}$, and thus belongs to an element of $\mathscr{B}$.

(ii) Let

$$B_1 = S_1 \cap \cdots \cap S_m, \quad B_2 = S_1' \cap \cdots \cap S_n'$$

be two elements of $\mathscr{B}$. Their intersection

$$B_1 \cap B_2 = (S_1 \cap \cdots \cap S_m) \cap (S_1' \cap \cdots \cap S_n')$$

is also a finite intersection of elements of $\mathscr{S}$, so it belongs to $\mathscr{B}$.

□

26.2 Examples of Topologies

Order Topology

Definition 26.12 (Order topology)**.** Let $(X,<)$, $|X|>1$. Let $\mathscr{B}$ be the collection of all sets of the following types:

(i) All open intervals (a,b) in X.

(ii) All intervals of the form $[a_0,b)$, where a_0 is the smallest element (if any) of X.

(iii) All intervals of the form $(a,b_0]$, where b_0 is the largest element (if any) of X.

The topology generated by $\mathscr{B}$ is called the ***order topology***.

We need to check that $\mathscr{B}$ is a basis of X.

(i) Every $x \in X$ lies in some element of $\mathscr{B}$: the smallest element (if any) lies in all sets of type (ii), the largest element (if any) lies in all sets of type (iii), and every other element lies in a set of type (i).

(ii) The intersection of any two sets of the preceding types is a set of one of these types, or is empty. Several cases need to be checked; we leave it to you.

For instance, let $x \in (a,b)\cap(c,d)$. Let $p=\max\{a,c\}$, $q=\min\{b,d\}$. Then $x \in (p,q) \subset (a,b)\cap(c,d)$, where $(p,q)\in\mathscr{B}$.

Example.

- The standard topology on $\mathbb{R}$ is just the order topology derived from the usual order on $\mathbb{R}$.

Definition 26.13. Let $(X,<)$, $a \in X$. Then the following subsets of X are ***rays*** determined by a:

$$\begin{aligned}(a,+\infty) &= \{x \in X \mid x > a\},\\ [a,+\infty) &= \{x \in X \mid x \geq a\},\\ (-\infty,a) &= \{x \in X \mid x < a\},\\ (-\infty,a] &= \{x \in X \mid x \leq a\}.\end{aligned}$$

$(a,+\infty)$ and $(-\infty,a)$ are called *open rays*, since they are open; for instance, $(a,+\infty)=\bigcup_{x>a}(a,x)$. Similarly, $[a,+\infty)$ and $(-\infty,a]$ are *closed rays*.

Lemma 26.14. *The collection of open rays form a subbasis for the order topology.*

Proof. Let $\mathscr{T}$ be the order topology on X, let $\mathscr{T}'$ be the topology generated by the subbasis of open rays. We will show that $\mathscr{T} = \mathscr{T}'$.

- Because the open rays are open in the order topology, the topology they generate is contained in the order topology. Hence $\mathscr{T}' \subset \mathscr{T}$.
- On the other hand, every basis element for the order topology equals a finite intersection of open rays; the interval (a,b) equals the intersection of $(-\infty,b)$ and $(a,+\infty)$, while $[a_0,b)$ and $(a,b_0]$, if they exist, are themselves open rays. Hence the topology generated by the open rays contains the order topology, so $\mathscr{T} \subset \mathscr{T}'$.

□

Product Topology

Let X and Y be topological spaces. We can define a topology on the cartesian product $X \times Y$.

Definition 26.15 (Product topology)**.** Let $(X, \mathscr{T}_X)$ and $(Y, \mathscr{T}_Y)$ be topological spaces. The ***product topology*** on $X \times Y$ is the topology $\mathscr{T}_{X\times Y}$ with basis

$$\mathscr{B} = \{U \times V \mid U \in \mathscr{T}_X, V \in \mathscr{T}_Y\}.$$

Remark. If $U \in \mathscr{T}_X$, $V \in \mathscr{T}_Y$, then $U \times V$ is an open set, but it need not be a basis element. **Counterexample:** Consider the union of two open rectangles in $\mathbb{R}^2$.

Hence the basis does not contain all the open subsets of $X \times Y$.

Lemma. *$\mathscr{B}$ is a basis.*

Proof.

(i) $X \times Y$ is a basis element, so every element of $X \times Y$ is contained in $X \times Y$.

(ii) Let $U_1 \times V_1, U_2 \times V_2 \in \mathscr{B}$. Then their intersection is

$$(U_1 \times V_1) \cap (U_2 \times V_2) = (U_1 \cap U_2) \times (V_1 \cap V_2).$$

Since $U_1 \cap U_2 \in \mathscr{T}_X$, $V_1 \cap V_2 \in \mathscr{T}_Y$, we have that $(U_1 \cap U_2) \times (V_1 \cap V_2) \in \mathscr{B}$.

□

What can one say if the topologies on X and Y are given by bases? The answer is as follows:

Lemma 26.16. *If $\mathscr{B}$ is a basis for the topology of X, $\mathscr{C}$ is a basis for the topology of Y, then the collection*

$$\mathscr{D} = \{B \times C \mid B \in \mathscr{B}, C \in \mathscr{C}\}$$

is a basis for the topology of $X \times Y$.

Proof. We apply 26.6.

Let W be an open set of $X \times Y$, and let $(x,y) \in W$. By definition of product topology there exists a basis element $U \times V$ such that $(x,y) \in U \times V \subset W$.

Since $\mathscr{B}$ and $\mathscr{C}$ are bases for X and Y respectively, there exists $B \in \mathscr{B}$ such that $x \in B \subset U$, and $C \in \mathscr{C}$ such that $y \in C \subset V$. Then $(x,y) \in B \times C \subset W$.

Thus the collection $\mathscr{D}$ meets the criterion of 26.6, so $\mathscr{D}$ is a basis for $X \times Y$. □

Definition 26.17 (Projection map)**.** Define the projection map of $X \times Y$ onto X as

$$\begin{aligned} \pi_1 : X \times Y &\to X \\ (x,y) &\mapsto x \end{aligned}$$

and the projection map of $X \times Y$ onto Y as

$$\begin{aligned} \pi_2 : X \times Y &\to Y \\ (x,y) &\mapsto y \end{aligned}$$

Remark. We use the word "onto" because π_1 and π_2 are surjective (unless one of the spaces X or Y happens to be empty, in which case $X \times Y$ is empty and our whole discussion is empty as well!).

If U is an open subset of X, then the set $\pi_1^{-1}(U)$ is precisely the set $U \times Y$, which is open in $X \times Y$. Similarly, if V is open in Y, then $\pi_2^{-1}(V) = X \times V$, which is open in $X \times Y$.

Lemma 26.18. *The collection*

$$\mathscr{S} = \{\pi_1^{-1}(U) \mid U \text{ open in } X\} \cup \{\pi_2^{-1}(V) \mid V \text{ open in } Y\}$$

is a subbasis for the product topology on $X \times Y$.

Proof. Let $\mathscr{T}$ denote the product topology on $X \times Y$; let $\mathscr{T}'$ be the topology generated by $\mathscr{S}$. We will show that $\mathscr{T} = \mathscr{T}'$.

$\boxed{\supset}$ Since every element of $\mathscr{S}$ belongs to $\mathscr{T}$, so do arbitrary unions of finite intersections of elements of $\mathscr{S}$. Thus $\mathscr{T}' \subset \mathscr{T}$.

$\boxed{\subset}$ Every basis element $U \times V$ for the topology $\mathscr{T}$ is a finite intersection of elements of $\mathscr{S}$, since

$$U \times V = \pi_1^{-1}(U) \cap \pi_2^{-1}(V).$$

Thus $U \times V$ belongs to $\mathscr{T}'$, so $\mathscr{T} \subset \mathscr{T}'$. □

Subspace Topology

Let X be a topological space. If $Y \subset X$, we can define a topology on Y as follows.

Definition 26.19 (Subspace topology)**.** Let $(X, \mathscr{T})$ be a topological space, and $Y \subset X$. The ***subspace topology*** on Y is the collection

$$\mathscr{T}_Y := \{Y \cap U \mid U \in \mathscr{T}\}.$$

With this topology, Y is called a ***subspace*** of X.

That is, the open sets of Y consist of all intersections of open sets of X with Y.

Lemma. *$\mathscr{T}_Y$ is a topology on Y.*

Proof.

(i) $\emptyset = Y \cap \emptyset$ so $\emptyset \in \mathscr{T}_Y$. $Y = Y \cap X$ so $Y \in \mathscr{T}_Y$.

(ii) $\mathscr{T}_Y$ is closed under finite intersections, since

$$(U_1 \cap Y) \cap \cdots \cap (U_n \cap Y) = (U_1 \cap \cdots \cap U_n) \cap Y.$$

(iii) $\mathscr{T}_Y$ is closed under arbitrary unions, since

$$\bigcup_{i \in I} (U_i \cap Y) = \left(\bigcup_{i \in I} U_i \right) \cap Y.$$

□

The next result provides the basis for the subspace topology.

Lemma 26.20. *If $\mathscr{B}$ is a basis for the topology of X, then*

$$\mathscr{B}_Y = \{B \cap Y \mid B \in \mathscr{B}\}$$

is a basis for the subspace topology on Y.

Proof. Let U be open in X, $y \in U \cap Y$. Since $\mathscr{B}$ is a basis for the topology of X, there exists $B \in \mathscr{B}$ such that $y \in B \subset U$. Then $y \in B \cap Y \subset U \cap Y$.

By 26.6, $\mathscr{B}_Y$ is a basis for the subspace topology on Y. □

When dealing with a space X and a subspace Y, one needs to be careful when one uses the term "open set". Does one mean an element of the topology of Y or an element of the topology of X?

We make the following definition: If Y is a subspace of X, we say that a set U is *open in Y* if it belongs to the topology of Y; this implies in particular that it is a subset of Y. We say that U is *open in X* if it belongs to the topology of X.

There is a special situation in which every set open in Y is also open in X:

Lemma 26.21. *Let Y be a subspace of X. If U is open in Y, and Y is open in X, then U is open in X.*

Proof. Since U is open in Y, $U = Y \cap V$ for some set V open in X. Since Y and V are both open in X, so is $Y \cap V$. □

Now let us explore the relation between the subspace topology and the product topology.

Lemma 26.22. *If A is a subspace of X, and B is a subspace of Y, then the product topology on $A \times B$ is the same as the topology $A \times B$ inherits as a subspace of $X \times Y$.*

Proof. The set $U \times V$ is the general basis element for $X \times Y$, where U is open in X and V is open in Y. Hence $(U \times V) \cap (A \times B)$ is the general basis element for the subspace topology on $A \times B$. Now

$$(U \times V) \cap (A \times B) = (U \cap A) \times (V \cap B).$$

Since $U \cap A$ and $V \cap B$ are the general open sets for the subspace topologies on A and B, respectively, the set $(U \cap A) \times (V \cap B)$ is the general basis element for the product topology on $A \times B$.

The conclusion we draw is that the bases for the subspace topology on $A \times B$ and for the product topology on $A \times B$ are the same. Hence the topologies are the same. □

26.3 Closed Sets and Limit Points

Let X be a topological space. If E is an open set containing x, we say E is a ***neighbourhood*** of x.

Closed Sets

Definition 26.23 (Closed set)**.** We say $E \subset X$ is ***closed*** if its complement $E^c = X \setminus E$ is open.

The collection of closed subsets of a space X has properties similar to those satisfied by the collection of open subsets of X:

Lemma 26.24. *Let X be a topological space.*

(i) $\emptyset$ and X are closed.

(ii) Arbitrary intersections of closed sets are closed.

(iii) Finite unions of closed sets are closed.

Proof.

(i) $\emptyset^c = X$ is open, so $\emptyset$ is closed. Similarly, $X^c = \emptyset$ is open, so X is closed.

(ii) Suppose $\{A_i\}_{i \in I}$ is a collection of closed sets. By de Morgan's laws,

$$\left(\bigcap_{i \in I} A_i \right)^c = \bigcup_{i \in I} {A_i}^c.$$

Since each ${A_i}^c$ is open, the RHS is open since it is an arbitrary union of open sets. Hence $\bigcap A_i$ is closed.

(iii) Suppose A_i is closed for $i = 1, \ldots, n$. Then

$$\left(\bigcup_{i=1}^{n} A_i \right)^c = \bigcap_{i=1}^{n} {A_i}^c.$$

The RHS is a finite intersection of open sets and is thus open. Hence $\bigcup A_i$ is closed.

□

If Y is a subspace of X, we say E is *closed in Y* if $E \subset Y$ and E is closed in the subspace topology of Y (that is, $Y \setminus E$ is open in Y).

Proposition 26.25. *Let Y be a subspace of X. Then E is closed in Y if and only if it equals the intersection of a closed set of X with Y.*

Proof.

$\boxed{\Longleftarrow}$ Suppose $E = C \cap Y$, where C is closed in X.

Then $X \setminus C$ is open in X, so that $(X \setminus C) \cap Y$ is open in Y, by definition of subspace topology.

But $(X \setminus C) \cap Y = Y \setminus E$. Hence $Y \setminus E$ is open in Y, so that E is closed in Y.

$\boxed{\Longrightarrow}$ Suppose E is closed in Y. Then $Y \setminus E$ is open in Y.

By definition of subspace topology, $Y \setminus E$ is the intersection of an open set U of X with Y. Note that $X \setminus U$ is closed in X, and $E = Y \cap (X \setminus U)$. Thus E equals the intersection of a closed set of X with Y, as desired. □

A set E that is closed in the subspace Y may or may not be closed in the larger space X. As was the case with open sets, there is a criterion for E to be closed in E:

Proposition 26.26. *Let Y be a subspace of X. If E is closed in Y, and Y is closed in X, then E is closed in X.*

Proof. □

Closure and Interior

Definition 26.27. Let $E \subset X$.
The ***interior*** E° of E is the union of all open sets contained in E.
The ***closure*** $\overline{E}$ of E is the intersection of all closed sets containing E.
The ***boundary*** of E is $\partial E := E \setminus E^\circ$.
We say E is ***dense*** in X if $\overline{E} = X$.

Hence the interior of a set is the *largest* open set contained in it; the closure of a set is the *smallest* closed set containing it.

Lemma 26.28.

(i) E *is open* $\iff E = E^\circ$.

(ii) E *is closed* $\iff E = \overline{E}$.

Lemma 26.29 (Properties). *Let $E, F \subset X$.*

(i) $(E \cap F)^\circ = E^\circ \cap F^\circ$, $(E \cup F)^\circ \supset E^\circ \cup F^\circ$.

(ii) $\overline{E \cup F} = \overline{E} \cup \overline{F}$, $\overline{E \cap F} \subset \overline{E} \cap \overline{F}$.

(iii) *If* $E \subset F$, *then* $\overline{E} \subset \overline{F}$.

When dealing with a topological space X and a subspace Y, one needs to exercise care in taking closures of sets. If E is a subset of Y, the closure of E in Y and the closure of E in X will in general be different.

Notation. We reserve the notation $\overline{E}$ to mean the closure of E in X.

The next result shows that the closure of E in Y can be expressed in terms of $\overline{E}$.

Proposition 26.30. *Let Y be a subspace of X; let $E \subset Y$, let $\overline{E}$ denote the closure of E in X. Then the closure of E in Y equals $\overline{E} \cap Y$.*

Proof. Let F denote the closure of E in Y. We will show that $F = \overline{E} \cap Y$.

$\boxed{\subset}$ $\overline{E}$ is closed in X, so $\overline{E} \cap Y$ is closed in Y.

Since $\overline{E} \cap Y$ contains E, and since by definition F equals the intersection of all closed subsets of Y containing E, we must have $F \subset (\overline{E} \cap Y)$.

$\boxed{\supset}$ Since F is closed in Y, $F = C \cap Y$ for some set C closed in X. Then C is a closed set of X containing E; since $\overline{E}$ is the intersection of all such closed sets, we have $\overline{E} \subset C$. Then $(\overline{E} \cap Y) \subset (C \cap Y) = F$. □

The next result provides a convenient characterisation of the closure of a set.

Lemma 26.31. *Let $E \subset X$. Then $x \in \overline{E}$ if and only if every neighbourhood of x intersects E.*

Proof.

$\boxed{\Longrightarrow}$ We prove the contrapositive. Suppose $x \notin \overline{E}$. Then $x \in \overline{E}^c$.

Thus $\overline{E}^c$ is an open set containing x which does not intersect E, as desired.

$\boxed{\Longleftarrow}$ We prove the contrapositive. Suppose there exists a neighbourhood U of x which does not intersect E. Then U^c is a closed set containing E. By definition of closure $\overline{E}$, U^c must contain $\overline{E}$; hence x cannot be in $\overline{E}$. □

Corollary 26.32. *Supposing the topology of X is given by a basis, then $x \in \overline{E}$ if and only if every basis element B containing x intersects E.*

Proof.

$\boxed{\Longrightarrow}$ If every open set containing x intersects E, so does every basis element B containing x, because B is an open set.

$\boxed{\Longleftarrow}$ If every basis element containing x intersects E, so does every open set U containing x, because U contains a basis element that contains x. □

Limit Points

Definition 26.33. Let $E \subset X$. We say $x \in X$ is a ***limit point*** of E, if every neighbourhood of x intersects E in some point other than x itself.

Let E' denote the set of all limit points of E.

We shall now see that limit points provide another way to describe the closure of a set.

Lemma 26.34. *Let $E \subset X$. Then $\overline{E} = E \cup E'$.*

Proof.

$\boxed{\supset}$ Let $x \in E'$. Then every neighbourhood of x intersects E (in a point different from x).

By 26.31, $x \in \overline{E}$. Hence $E' \subset \overline{E}$. Since $E \subset \overline{E}$, it follows that $E \cup E' \subset \overline{E}$.

$\boxed{\subset}$ Let $x \in \overline{E}$.

- If $x \in E$, it is trivial that $x \in E \cup E'$.
- If $x \notin E$, since $x \in \overline{E}$, we know that every neighbourhood U of x intersects A. Since $x \notin E$, the set U must intersect E in a point different from x. Then $x \in E'$.

Thus $x \in E \cup E'$, so $\overline{E} \subset E \cup E'$. □

Corollary 26.35. *$E \subset X$ is closed if and only if it contains all its limit points.*

Proof. E is closed if and only if $E = \overline{E}$, and the latter holds if and only if $E' \subset E$. □

Hausdorff Spaces

Let X be a topological space.

Definition 26.36 (Hausdorff space)**.** We say X is a ***Hausdorff space*** if, for all distinct $x, y \in X$, there exist neighbourhoods U and V of x and y respectively that are disjoint.

Lemma 26.37. *Every finite point set in a Hausdorff space X is closed.*

Proof. It suffices to show that every one-point set $\{x_0\}$ is closed.

If x is a point of X different from x_0, then x and x_0 have disjoint neighbourhoods U and V, respectively. Since U does not intersect $\{x_0\}$, the point x cannot belong to the closure of the set $\{x_0\}$.

Hence the closure of $\{x_0\}$ is $\{x_0\}$ itself, so $\{x_0\}$ is closed. □

The condition that finite point sets be closed is called the T_1 ***axiom***:

> For all distinct $p,q \in X$, there exists neighbourhoods U,V which contain p,q respectively but not the other.

(Note that U and V can intersect.)

Proposition 26.38. *Let X be a space satisfying the T_1 axiom; let $E \subset X$. Then x is a limit point of E if and only if every neighbourhood of x contains infinitely many points of E.*

Proof.

$\boxed{\Longleftarrow}$ If every neighbourhood of x intersects E in infinitely many points, it certainly intersects E in some point other than x itself, so that x is a limit point of E.

$\boxed{\Longrightarrow}$ Let x be a limit point of E. Suppose, for a contradiction, that there exists a neighbourhood U of x which intersects E at finitely many points.

Then U also intersects $E \setminus \{x\}$ at finitely many points, say $x_1,\ldots,x_m$; that is,

$$U \cap (E \setminus \{x\}) = \{x_1,\ldots,x_m\}.$$

Since the finite point set $\{x_1,\ldots,x_m\}$ is closed, its complement $\{x_1,\ldots,x_m\}^c$ is open. Then

$$U \cap \{x_1,\ldots,x_m\}^c$$

is a neighbourhood of x that does not intersect $E \setminus \{x\}$. This contradicts the assumption that x is a limit point of E. □

Definition 26.39 (Limit). If a sequence (x_n) in a Hausdorff space X converges to $x \in X$, we write $x_n \to x$, and say x is the ***limit*** of (x_n).

Lemma 26.40 (Uniqueness of limit). *If X is a Hausdorff space, then a sequence of points of X converges to at most one point of X.*

Proof. Suppose (x_n) is a sequence in X, $x_n \to x$.

If $y \neq x$, let U and V be disjoint neighbourhoods of x and y, respectively. Since U contains x_n for all but finitely many values of n, the set V cannot. Therefore, $x_n \not\to y$. □

Proposition 26.41. *Every simply ordered set is a Hausdorff space in the order topology. The product of two Hausdorff spaces is a Hausdorff space. A subspace of a Hausdorff space is a Hausdorff space.*

26.4 Continuous Functions

Continuity of a Function

Let X and Y be topological spaces.

Definition 26.42. We say $f\colon X \to Y$ is ***continuous*** if for every open set $V \subset Y$, $f^{-1}(V)$ is open in X.

That is, the pre-images of open sets are open.

Remark. Continuity of a function depends not only upon the function f itself, but also on the topologies specified for its domain and range. If we wish to emphasise this fact, we can say that f is continuous *relative* to specific topologies on X and Y.

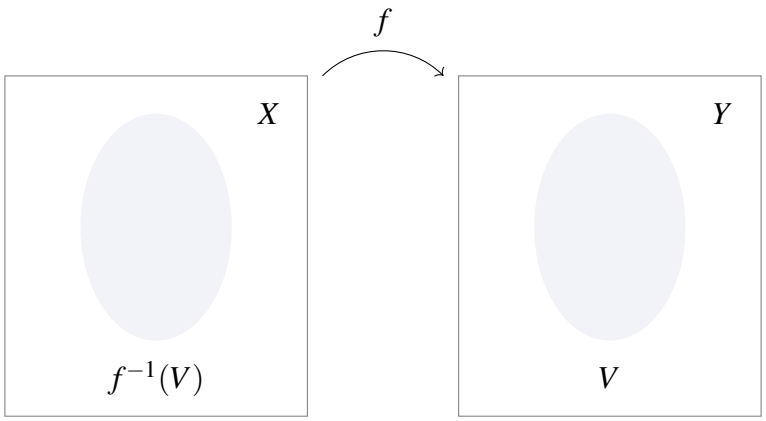


Figure 26.3: Pre-image of a set

The definition of continuity in Definition 26.42 is a global one. Frequently it is desirable to define continuity *locally* at each point of X:

> We say $f\colon X \to Y$ is *continuous* at $x \in X$ if, for every neighbourhood V of $f(x)$, there exists a neighbourhood U of x such that $f(U) \subset V$.

When X and Y are metric spaces, the local and global definitions are equivalent. The following characterisation of continuous functions shows that the local and global definitions of continuity are equivalent for topological spaces.

Lemma 26.43. *Let $f\colon X \to Y$. Then the following are equivalent:*

(i) f is continuous.

(ii) f is continuous at every $x \in X$.

(iii) For every $A \subset X$, $f(\overline{A}) \subset \overline{f(A)}$.

(iv) For every $B \subset Y$, $\overline{f^{-1}(B)} \subset f^{-1}(\overline{B})$.

(v) For every closed set $B \subset Y$, $f^{-1}(B)$ is closed in X.

Proof.

$\boxed{\text{(i)} \implies \text{(ii)}}$ Let $x \in X$ and let V be a neighbourhood of $f(x)$. Then the set $U = f^{-1}(V)$ is a neighbourhood of x such that $f(U) \subset V$.

$\boxed{\text{(ii)} \implies \text{(i)}}$ Let V be an open set of Y; let x be a point of $f^{-1}(V)$. Then $f(x) \in V$, so that by hypothesis there exists a neighbourhood U_x of x such that $f(U_x) \subset V$. Then $U_x \subset f^{-1}(V)$. It follows that $f^{-1}(V)$ can be written as the union of the open sets U_x, so that it is open.

$\boxed{\text{(i)} \implies \text{(iii)}}$ Suppose f is continuous. Let $A \subset X$. We show that if $x \in \overline{A}$, then $f(x) \in \overline{f(A)}$. Let V be a neighbourhood of $f(x)$. Then $f^{-1}(V)$ is an open set of X containing x; it must intersect A in some point y. Then V intersects $f(A)$ in the point $f(y)$, so that $f(x) \in \overline{f(A)}$, as desired. □

The next result states that continuous functions of continuous functions are continuous.

Lemma 26.44. *Let X, Y and Z be topological spaces. If $f\colon X \to Y$ and $g\colon Y \to Z$ are continuous, then $h = g \circ f$ is continuous.*

Proof. Let U be open in Z. By continuity of g, we have that $g^{-1}(U)$ is open in Y. Note that

$$h^{-1}(U) = f^{-1}\left(g^{-1}(U)\right).$$

If f is continuous, it follows that $h^{-1}(U)$ is open, so h is continuous. □

Homeomorphisms

Definition 26.45 (Homeomorphism)**.** Let X and Y be topological spaces; let $f\colon X \to Y$ be a bijection. We say f is a ***homeomorphism*** if f and f^{-1} are continuous.
In this case, we say X and Y are *homeomorphic*.

Constructing Continuous Functions

26.5 Metric Topology

One of the most important and frequently used ways of imposing a topology on a set is to define the topology in terms of a metric on the set.

Definition 26.46.

26.6 Quotient Topology

Definition 26.47 (Quotient map)**.** Let X and Y be topological spaces. Let $p: X \to Y$ be surjective. We say p is a ***quotient map*** if

$$U \text{ is open in } Y \iff p^{-1}(U) \text{ is open in } X.$$

Remark. This condition is stronger than continuity.

Instead of defining quotient maps using open sets, we can use closed sets:

Lemma 26.48. *Let $p: X \to Y$ be surjective. Then p is a quotient map if and only if*

$$A \text{ is closed in } Y \iff p^{-1}(A) \text{ is closed in } X.$$

Proof. Equivalence of the two conditions follows from $p^{-1}(U^c) = [p^{-1}(U)]^c$. □

Let $p: X \to Y$ be surjective. We say $C \subset X$ is *saturated* (with respect to p) if

$$\forall y \in Y, \quad p^{-1}(\{y\}) \cap C \neq \emptyset \implies p^{-1}(\{y\}) \subset C.$$

That is, C is saturated $\iff C = p^{-1}(A)$ for some $A \subset Y$.

Lemma 26.49. *Let $p: X \to Y$ be surjective. Then p is a quotient map if and only if p is continuous and p maps saturated open sets of X to open sets of Y.*

Proof.

$\boxed{\implies}$ Let $U \subset Y$ be open and saturated. Then there exists $A \subset Y$ such that $U = p^{-1}(A)$. Since U is open, $p^{-1}(A)$ is open in X. Since p is a quotient map, this implies $A = p(U)$ is open in Y.

$\boxed{\impliedby}$ Suppose $p^{-1}(U)$ is open in X. □

Two special kinds of quotient maps are the open maps and the closed maps.

- We say $f: X \to Y$ is an *open map* if for each open set U of X, $f(U)$ is open in Y.
- We say $f: X \to Y$ is a *closed map* if for each closed set A of X, $f(A)$ is closed in Y.

We have a sufficient condition for quotient maps:

Lemma 26.50. *If $p: X \to Y$ is a surjective continuous map that is either open or closed, then p is a quotient map.*

Proof. This follows immediately from the definition. □

Definition 26.51 (Quotient topology)**.** Let A be any set. If $p\colon X \to A$ is surjective, there exists a unique topology $\mathscr{T}$ on A, relative to which p is a quotient map.
We call $\mathscr{T}$ the ***quotient topology*** induced by p.

Lemma. *The quotient topology $\mathscr{T}$ is a topology.*

Definition 26.52 (Quotient space)**.** Let X be a topological space, and let X^* be a partition of X into disjoint subsets whose union is X. Let $p\colon X \to X^*$ be the surjective map that carries each point of X to the element of X^* containing it. In the quotient topology induced by p, the space X^* is called a ***quotient space*** of X.

27 Connectedness and Compactness

27.1 Connected Spaces

The definition of connectedness for a topological space is a quite natural one. One says that a space can be "separated" if it can be broken up into two "globs" – disjoint open sets; otherwise, one says that it is connected. From this simple idea much follows.

Let X be a topological space.

Definition 27.1 (Connectedness). A *separation* of X is a pair U, V of disjoint non-empty open subsets of X with $X = U \cup V$.
We say X is ***connected*** if there does not exist a separation of X.

Remark. The definition of connectedness is a negation statement. This makes many arguments involving connectedness naturally take the form of proof by *contradiction*: one assumes the existence of a separation, and then deduces an inconsistency.

Another way of formulating the definition of connectedness is the following:

Lemma 27.2. *A space X is connected if and only if the only subsets of X that are both open and closed are $\emptyset$ and X itself.*

Proof.

$\boxed{\Longrightarrow}$ Suppose, for a contradiction, that there exists a non-empty proper subset A of X that is both open and closed in X. Then the sets A and A^c constitute a separation of X, for they are open, disjoint, and non-empty, and their union is X.

This contradicts the connectedness of X.

$\boxed{\Longleftarrow}$ If U and V form a separation of X, then U is non-empty and different from X, and it is both open and closed in X. □

For a subspace Y of a topological space X, there is another useful way of formulating the definition of connectedness:

Lemma 27.3. *If Y is a subspace of X, a separation of Y is a pair of disjoint non-empty sets A and B whose union is Y, neither of which contains a limit point of the other. The space Y is connected if there exists no separation of Y.*

Lemma 27.4. *If C and D form a separation of X, and if Y is a connected subspace of X, then $Y \subset C$ or $Y \subset D$.*

Proof. Since C and D are open in X, the sets $Y \cap C$ and $Y \cap D$ are open in Y. These two sets are disjoint (since C and D are disjoint), and their union is

$$(Y \cap C) \cup (Y \cap D) = Y \cap (C \cup D) = Y \cap X = Y.$$

If they were both non-empty, they would constitute a separation of Y (which contradicts the connectedness of Y). Therefore, one of them is empty. Hence Y must lie entirely in C or in D. □

Proposition 27.5. *An arbitrary union of a connected subspaces of X that have a point in common is connected.*

Proof. prove by contradiction □

Lemma 27.6. *If C and D are open and disjoint, then $\overline{C} \cap D = \emptyset$.*

Proof. We prove by contradiction. Let $x \in \overline{C} \cap D$. Then $x \in D$, so $x \notin C$. Since $x \in \overline{C}$, $x \in \partial C$.

Since D is open, there is a neighbourhood of x contained in D. But this neighbourhood must intersect C, since $x \in \partial C$. This contradicts $C \cap D = \emptyset$. □

Proposition 27.7. *Let A be a connected subspace of X. If $A \subset B \subset \overline{A}$, then B is also connected.*

Proof. Suppose, for a contradiction, that B is not connectedness. Then $B = C \cup D$ is a separation of B.

By 27.3, either $A \subset C$ or $A \subset D$; WLOG assume $A \subset C$. Then $\overline{A} \subset \overline{C}$; since $\overline{C}$ and D are disjoint, B cannot intersect D. This contradicts the fact that D is a non-empty subset of B. □

Proposition 27.8. *The image of a connected subspace under a continuous map is connected.*

Proof. Suppose $f\colon X \to Y$ is continuous, and $A \subset X$ is a connected subspace. We will show $f(A)$ is connected.

□

27.2 Connected Subspaces of $\mathbb{R}$

27.3 Components and Local Connectedness

27.4 Compact Spaces

Definition 27.9. Let X be a space, $K \subset X$. We say a collection of open subsets $\{U_i\}_{i\in I}$ is an *open cover* of K if

$$K \subset \bigcup_{i\in I} U_i.$$

A *finite subcover* is a finite subcollection of $\{U_i\}$ which covers K.

Definition 27.10 (Compactness)**.** We say $K \subset X$ is ***compact*** if every open cover of X contains a finite subcover.
In particular, if X is itself compact, we say X is a *compact space*.

Example. Any finite set is compact.

Proof. Let $S = \{x_1, \ldots, x_n\}$ be a finite set. Let $\{U_i\}_{i\in I}$ be an open cover of S. For each x_k, choose *one* U_{i_k} such that $x_k \in U_{i_k}$. Then $\{U_{i_1}, \ldots, U_{i_n}\}$ is a finite subcover. □

Lemma 27.11. *Let Y be a subspace of X. Then Y is compact if and only if every open cover of Y (consisting of open sets in X) contains a finite subcover.*

Proof.

$\boxed{\Longrightarrow}$ Suppose Y is comapct, and $\{U_i\}$ is an open cover of Y, where U_i are open in X. Then the collection

$$\{U_i \cap Y \mid i \in I\}$$

is an open cover of Y, where $U_i \cap Y$ are open in Y. Hence a finite subcollection

$$\{U_{i_1} \cap Y, \ldots, U_{i_n} \cap Y\}$$

covers Y; thus it is a finite subcover of Y.

$\boxed{\Longleftarrow}$ Suppose the given condition holds; we wish to prove Y compact.

Let □

The next result states that a closed subset of a compact set is compact.

Lemma 27.12. *Suppose K is compact and F is closed, in a topological space X. If $F \subset K$, then F is compact.*

Proof. Let $\{U_i\}_{i\in I}$ be an open cover of F. Since F is closed, F^c is open; let $W = F^c$, then

$W \cup \bigcup_{i\in I} U_i$ is an open cover of X. Hence there exists a finite subcover $\{U_{i_k}\}$ such that

$$K \subset \bigcup_{k=1}^{n} U_{i_k}.$$

Then $F \subset \bigcup_{k=1}^{n} U_{i_k}$. □

Corollary 27.13. *Suppose $A \subset B$. If $\overline{B}$ is compact, then $\overline{A}$ is compact.*

Proposition 27.14. *Suppose X is a Hausdorff space, $K \subset X$ is compact, and $x \in K^c$. Then there exist open sets U and W such that $x \in U$, $K \subset W$, and $U \cap W = \emptyset$.*

Proof. Let $y \in K$. The Hausdorff separation axiom implies the existence of disjoint open sets U_y and V_y such that $x \in U_y$ and $y \in U_y$. Since K is compact, there exist points $y_1, \ldots, y_n \in K$ such that

$$K \subset V_{y_1} \cup \cdots \cup V_{y_n}.$$

Our requirements are then satisfied by the sets

$$U = U_{y_1} \cap \cdots \cap U_{y_n} \quad \text{and} \quad W = V_{y_1} \cup \cdots \cup V_{y_n}.$$

□

Corollary 27.15. *Compact subsets of Hausdorff spaces are closed.*

Corollary 27.16. *If F is closed and K is compact in a Hausdorff space, then $F \cap K$ is compact.*

Proof. This follows from the previous result, and closed subset of compact set is compact. □

Proposition 27.17. *Let $\{K_i\}_{i\in I}$ be a collection of compact subsets of a Hausdorff space. If $\bigcap_{i\in I} K_i = \emptyset$, then some finite subcollection of $\{K_i\}$ also has empty intersection.*

Proof. Let $V_i = K_i{}^c$. Fix a member K_1 of $\{K_i\}$. Since no point of K_1 belongs to every K_i, $\{V_i\}$ is an open cover of K_1. Hence there exist $i_1, \ldots, i_n$ such that

$$K_1 \subset V_{i_1} \cup \cdots \cup V_{i_n}.$$

This implies that

$$K_1 \cap K_{i_1} \cap \cdots \cap K_{i_n} = \emptyset.$$

□

Proposition 27.18. *Let $f: X \to Y$ be continuous, and $A \subset X$ is compact. Then $f(A)$ is compact (in Y).*

Proof. Let $\{U_i\}_{i\in I}$ be an open cover of $f(A)$ by sets open in Y. Consider the collection

$$\{f^{-1}(U_i)\}_{i\in I}.$$

These sets are open in X because f is continuous. Then

$$A \subset f^{-1}(f(A)) \subset f^{-1}\left(\bigcup_{i\in I} U_i\right) = \bigcup_{i\in I} f^{-1}(U_i)$$

so $\{f^{-1}(U_i)\}_{i\in I}$ is an open cover of A. Since A is compact, there exists $i_1, \ldots, i_n$ such that

$$A \subset \bigcup_{k=1}^{n} f^{-1}(U_{i_k}).$$

Then

$$f(A) \subset \bigcup_{k=1}^{n}$$

□

There is one final criterion for a space to be compact, a criterion that is formulated in terms of closed sets rather than open sets. First we make a definition.

Definition 27.19. A collection $\mathscr{U}$ of subsets of X is said to have the ***finite intersection property*** if every finite subcollection has a non-empty intersection.

Theorem 27.20. *Let X be a topological space. Then X is compact if and only if for every collection $\mathscr{C}$ of closed sets in X having the finite intersection property, $\bigcap_{C\in\mathscr{C}} \neq \emptyset$.*

A special case of this theorem occurs when we have a *nested sequence* $C_1 \supset \cdots \supset C_n \supset \cdots$ of closed sets in a compact space X. If each of the sets C_n is non-empty, then the collection $\mathscr{C} = \{C_n\}_{n\in\mathbb{N}}$ automatically has the finite intersection property. Then the intersection

$$\bigcap_{n=1}^{\infty} C_n$$

is non-empty.

27.5 Compact Subspaces of $\mathbb{R}$

27.6 Limit Point Compactness

27.7 Local Compactness

X is *locally compact* if every point of X has a neighbourhood whose closure is compact.

Obviously, every compact space is locally compact.

We recall the Heine-Borel theorem: The compact subsets of a euclidean space $\mathbb{R}^n$ are precisely those that are closed and bounded. From this it follows easily that $\mathbb{R}^n$ is a locally compact Hausdorff space.

Proposition 27.21. *Suppose U is open in a locally compact Hausdorff space X, $K \subset U$, and K is compact. Then there exists an open set V with compact closure such that*

$$K \subset V \subset \overline{V} \subset U.$$

Proof. Since every point of K has a neighborhood with compact closure, and since K is covered by the union of finitely many of these neighborhoods, K lies in an open set G with compact closure. If $U = X$, take $V = G$.

Otherwise, let C be the complement of U. Theorem 2.5 shows that to each $p \in C$ there corresponds an open set W_p such that $K \subset W_p$ and $p \notin \overline{W}_p$. Hence $\left\{C \cap \overline{G} \cap \overline{W}_p\right\}$, where p ranges over C, is a collection of compact sets with empty intersection. By Theorem 2.6 there exist points $p_1, \ldots, p_n \in C$ such that

$$C \cap \overline{G} \cap \overline{W}_{p_1} \cap \cdots \cap \overline{W}_{p_n} = \emptyset.$$

The set

$$V = G \cap W_{p_1} \cap \cdots \cap W_{p_n}$$

then has the required properties, since

$$\overline{V} \subset \overline{G} \cap \overline{W}_{p_1} \cap \cdots \cap \overline{W}_{p_n}.$$

□

VIII
Measure Theory

In measure theory, the main idea is that we want to assign "sizes" to different sets. For example, we might think $[0,2] \subset \mathbb{R}$ has size 2, while perhaps $\mathbb{Q} \subset \mathbb{R}$ has size 0. This is known as a *measure*.

One of the main applications of a measure is that we can use it to come up with a new definition of an integral, known as the *Lebesgue integral*. Instead of integrating functions $[a,b] \to \mathbb{R}$ only, we can replace the domain with any measure space, allowing us to integrate a much wider class of functions.

28 Measures

28.1 Introduction

One of the most venerable problems in geometry is to determine the area or volume of a region in the plane or in 3-space. The techniques of integral calculus provide a satisfactory solution to this problem for regions that are bounded by "nice" curves or surfaces but are inadequate to handle more complicated sets, even in dimension one.

Ideally, for $n \in \mathbb{N}$ we would like to have a function $\mu\colon \mathscr{P}(\mathbb{R}^n) \to [0,\infty]$ that assigns to each $E \subset \mathbb{R}^n$ a number $\mu(E) \in [0,\infty]$, the n-dimensional measure of E. We would desire μ to possess the following properties:

(i) If $E_1, E_2, \ldots$, is a finite or infinite sequence of disjoint sets, then

$$\mu(E_1 \cup E_2 \cup \cdots) = \mu(E_1) + \mu(E_2) + \cdots.$$

(ii) If E is congruent to F (that is, if E can be transformed into F by translations, rotations, and reflections), then $\mu(E) = \mu(F)$.

(iii) $\mu(Q) = 1$, where Q is the unit cube

$$Q = \{x \in \mathbb{R}^n \mid 0 \le x_i < 1,\ i = 1, \ldots, n\}.$$

Unfortunately, these conditions are mutually inconsistent.

Proposition 28.1. *There does not exist $\mu\colon \mathscr{P}(\mathbb{R}^n) \to [0,\infty]$ which satisfies (i), (ii), and (iii).*

Proof. It suffices to prove the case when $n = 1$.

Define an equivalence relation on $[0,1)$ by

$$x \sim y \iff x - y \in \mathbb{Q}.$$

Let N be a subset of $[0,1)$ that contains precisely one member of each equivalence class of $\sim$. (To find such an N, one must invoke the axiom of choice.)

Let $R = \mathbb{Q} \cap [0,1)$. For each $r \in R$, let

$$\begin{aligned} N_r &= ((N+r) \cap [0,1)) \cup ((N+r-1) \cap [0,1)) \\ &= \{x+r \mid x \in N \cap [0,1-r)\} \cup \{x+r-1 \mid x \in N \cap [1-r,1)\}. \end{aligned}$$

That is, to obtain N_r, shift N to the right by r units and then shift the part that sticks out beyond $[0,1)$ one unit to the left. Then $N_r \subset [0,1)$.

Lemma. *Every $x \in [0,1)$ belongs to precisely one N_r.*

Let y be the element of N that belongs to the equivalence class of x. Then $x \in N_r$ where $r = x-y$ if $x \geq y$, or $r = x-y+1$ if $x < y$.

On the other hand, let $x \in N_r \cap N_s$. Then $x-r$ (or $x-r+1$) and $x-s$ (or $x-s+1$) would be distinct elements of N belonging to the same equivalence class, which is impossible.

Suppose $\mu : \mathscr{P}(\mathbb{R}) \to [0,\infty]$ satisfies (i), (ii), and (iii). By (i) and (ii),

$$\mu(N) = \mu(N \cap [0,1-r)) + \mu(N \cap [1-r,1)) = \mu(N_r)$$

for any $r \in R$. Also, since R is countable and $[0,1)$ is the disjoint union of the N_r's,

$$\mu([0,1)) = \sum_{r \in R} \mu(N_r)$$

by (i) again. But $\mu([0,1)) = 1$ by (iii), and since $\mu(N_r) = \mu(N)$, the sum on the right is either 0 (if $\mu(N) = 0$) or ∞ (if $\mu(N) > 0$). Hence no such μ can exist. □

Faced with this discouraging situation, one might consider weakening (i) so that additivity is required to hold only for finite sequences. This is not a very good idea, as we shall see: The additivity for countable sequences is what makes all the limit and continuity results of the theory work smoothly. Moreover, in dimensions $n \geq 3$, even this weak form of (i) is inconsistent with (ii) and (iii). Indeed, in 1924 Banach and Tarski proved the following amazing result:

Theorem 28.2 (Banach–Tarski paradox). *Let $n \geq 3$. Let $A, B \subset \mathbb{R}^n$ be bounded and open. Then there exist partitions of A and B into a finite number of disjoint subsets, $A = A_1 \cup \cdots \cup A_k$, $B = B_1 \cup \cdots \cup B_k$ (for some integer k), such that A_i is congruent to B_i for $i = 1, \ldots, k$.*

Thus one can cut up a ball the size of a pea into a finite number of pieces and rearrange them to form a ball the size of the earth! Needless to say, the existence of the sets A_i and B_i clearly precludes the construction of any $\mu : \mathscr{P}(\mathbb{R}^n) \to [0,\infty]$ that assigns positive, finite values to bounded open sets and satisfies (i) for finite sequences as well as (ii).

The moral of these examples is that $\mathbb{R}^n$ contains subsets which are so strangely put together that it is impossible to define a geometrically reasonable notion of measure for them, and the remedy for the situation is to discard the requirement that μ should be defined on all subsets of $\mathbb{R}^n$. Rather, we shall content ourselves with constructing μ on a class of subsets of $\mathbb{R}^n$. To generalise, we shall deal with arbitrary sets X instead of $\mathbb{R}^n$.

28.2 σ-Algebras

In this section we discuss the families of sets that serve as the domains of measures.

Definitions and Properties

Definition 28.3 (σ-algebra)**.** Let X be a non-empty set. We say a non-empty collection of subsets $\mathscr{A} \subset \mathscr{P}(X)$ is an ***σ-algebra*** on X if

(i) $\emptyset, X \in \mathscr{A}$;

(ii) if $E_1, E_2, \cdots \in \mathscr{A}$, then $\bigcup_{n=1}^{\infty} E_n \in \mathscr{A}$; (closed under countable unions)

(iii) if $E \in \mathscr{A}$, then $E^c \in \mathscr{A}$. (closed under complements)

We call $(X, \mathscr{A})$ a ***measurable space***. The members of $\mathscr{A}$ are called *measurable sets* in X.

Notation. If the σ-algebra $\mathscr{A}$ is clear, we simply omit it and denote a measurable space as X.

If (ii) is closed under finite unions instead, we call $\mathscr{A}$ an *algebra* on X.

Suppose $\mathscr{A}$ is an algebra of subsets of X. Then we immediately deduce the following properties.

- If $E_1, \ldots, E_n \in \mathscr{A}$, by de Morgan's laws,

$$\bigcap_{i=1}^{n} E_i = \left(\bigcup_{i=1}^{n} E_i^c \right)^c \in \mathscr{A}$$

  so $\mathscr{A}$ is closed under finite intersections. If $\mathscr{A}$ is a σ-algebra, then $\mathscr{A}$ is closed under countable intersections.

- If $A, B \in \mathscr{A}$, since $A \setminus B = B^c \cap A$, then $A \setminus B \in \mathscr{A}$.

Example.

- If X is any set, $\mathscr{P}(X)$ is a σ-algebra.
- If X is any set, $\{\emptyset, X\}$ is a σ-algebra.
- If X is any set, and $E \subset X$, $\{\emptyset, E, E^c, X\}$ is a σ-algebra.
- If X is uncountable, then

$$\mathscr{A} = \{E \subset X \mid E \text{ or } E^c \text{ is countable}\}$$

  is a σ-algebra, called the *σ-algebra of countable or co-countable sets*.

Borel σ-Algebra

The next result guarantees that there is a *smallest* σ-algebra on a set X containing a given set $\mathscr{E}$ of subsets of X.

Lemma 28.4. *Suppose X is a set, and $\mathscr{E} \subset \mathscr{P}(X)$. Then the intersection of all σ-algebras on X that contain $\mathscr{E}$*

$$\mathscr{M}(\mathscr{E}) = \bigcap_{\substack{\mathscr{A} \supset \mathscr{E} \\ \mathscr{A} \text{ is } \sigma\text{-algebra}}} \mathscr{A}.$$

is a σ-algebra on X.

We call $\mathscr{M}(\mathscr{E})$ the σ-algebra *generated* by $\mathscr{E}$.

Proof. $\mathscr{M}(\mathscr{E})$ is non-empty, since $\mathscr{P}(X)$ is a σ-algebra that contains $\mathscr{E}$. We now show $\mathscr{M}(\mathscr{E})$ is a σ-algebra on X:

(i) For every σ-algebra $\mathscr{A}$ that contains $\mathscr{E}$, $\emptyset \in \mathscr{A}$. Hence $\emptyset \in \mathscr{M}(\mathscr{E})$.

(ii) Let $E \in \mathscr{M}(\mathscr{E})$. Then E is in every σ-algebra on X that contains $\mathscr{E}$. Thus E^c is in every σ-algebra on X that contains $\mathscr{E}$. Hence $E^c \in \mathscr{M}(\mathscr{E})$.

(iii) Let (E_n) be a sequence of elements of $\mathscr{M}(\mathscr{E})$. Thus each E_n is in every σ-algebra on X that contains $\mathscr{E}$. Thus $\bigcup_{n=1}^{\infty} E_n$ is in every σ-algebra on X that contains $\mathscr{E}$. Hence $\bigcup_{n=1}^{\infty} E_n \in \mathscr{M}(\mathscr{E})$.

□

The following observation is often useful:

Lemma 28.5. *If $\mathscr{E} \subset \mathscr{F}$, then $\mathscr{M}(\mathscr{E}) \subset \mathscr{M}(\mathscr{F})$.*

Proof. $\mathscr{F}$ is a σ-algebra containing $\mathscr{E}$; it therefore contains $\mathscr{M}(\mathscr{E})$. □

We have come to an important example of a σ-algebra.

Definition 28.6 (Borel σ-algebra)**.** Let X be a topological space. Let $\mathscr{T}_X$ denote the collection of open sets in X. We define the ***Borel σ-algebra*** on X as

$$\mathscr{B}(X) := \mathscr{M}(\mathscr{T}_X).$$

The members of a Borel σ-algebra are called *Borel sets*.
A ***Borel measurable space*** is a set together with a Borel σ-algebra on it.

Hence $\mathscr{B}(X)$ contains open sets, closed sets, countable intersections of open sets, countable unions of closed sets, and so forth. We introduce some standard terminology for the levels in this hierarchy. [1]

- A countable intersection of open sets is called a G_δ set.
- A countable union of closed sets is called a F_σ set.
- A countable union of G_δ sets is called a $G_{\delta\sigma}$ set; a countable intersection of F_σ sets is called a $F_{\sigma\delta}$ set; and so forth.

The Borel σ-algebra on $\mathbb{R}$ will play a fundamental role in what follows. For future reference we note that it can be generated in a number of different ways:

Proposition 28.7. *$\mathscr{B}(\mathbb{R})$ is generated by each of the following:*

(i) the open intervals: $\mathscr{E}_1 = \{(a,b) \mid a < b\}$,

(ii) the closed intervals: $\mathscr{E}_2 = \{[a,b] \mid a < b\}$,

(iii) the half-open intervals: $\mathscr{E}_3 = \{(a,b] \mid a < b\}$ or $\mathscr{E}_4 = \{[a,b) \mid a < b\}$,

(iv) the open rays: $\mathscr{E}_5 = \{(a,\infty) \mid a \in \mathbb{R}\}$ or $\mathscr{E}_6 = \{(-\infty,a) \mid a \in \mathbb{R}\}$,

(v) the closed rays: $\mathscr{E}_7 = \{[a,\infty) \mid a \in \mathbb{R}\}$ or $\mathscr{E}_8 = \{(-\infty,a] \mid a \in \mathbb{R}\}$.

Proof. We want to show $\mathscr{M}(\mathscr{E}_i) = \mathscr{B}(\mathbb{R})$ for all i.

$\boxed{\subset}$ The elements of $\mathscr{E}_i$ for $i \neq 3,4$ are open or closed, so $\mathscr{E}_i \subset \mathscr{B}(\mathbb{R})$ for $i \neq 3,4$.

The elements of $\mathscr{E}_3$ and $\mathscr{E}_4$ are G_δ sets:

$$(a,b] = \bigcap_{n=1}^{\infty} \left(a, b + \frac{1}{n}\right), \quad [a,b) = \bigcap_{n=1}^{\infty} \left(a + \frac{1}{n}, b\right).$$

All of these are Borel sets. Thus $\mathscr{E}_i \subset \mathscr{B}(\mathbb{R})$.

Hence by 28.5, $\mathscr{M}(\mathscr{E}_i) \subset \mathscr{B}(\mathbb{R})$ for all i.

$\boxed{\supset}$ Every open set in $\mathbb{R}$ is a countable union of open intervals, so $\mathscr{T}_{\mathbb{R}} \subset \mathscr{E}_1$. By 28.5, $\mathscr{B}(\mathbb{R}) \subset \mathscr{M}(\mathscr{E}_1)$.

That $\mathscr{B}(\mathbb{R}) \subset \mathscr{M}(\mathscr{E}_i)$ for $i \geq 2$ can now be established by showing that all open intervals lie in

[1] δ and σ stand for the German "Durchschnitt" and "Summe", that is, intersection and union.

$\mathscr{M}(\mathscr{E}_i)$, and then apply 28.5:

$$(a,b) = \bigcup_{n=1}^{\infty} \left[a+\frac{1}{n}, b-\frac{1}{n}\right] \in \mathscr{M}(\mathscr{E}_2)$$
$$(a,b) = \bigcup_{n=1}^{\infty} (a, b-\frac{1}{n}] \in \mathscr{M}(\mathscr{E}_3)$$
$$(a,b) = \bigcup_{n=1}^{\infty} [a+\frac{1}{n}, b) \in \mathscr{M}(\mathscr{E}_4)$$
$$(a,b] = (a,\infty) \cap (b,\infty)^c \in \mathscr{M}(\mathscr{E}_5) \implies \mathscr{E}_3 \subset \mathscr{M}(\mathscr{E}_5)$$
$$[a,b) = (-\infty,a)^c \cap (b,\infty) \in \mathscr{M}(\mathscr{E}_6) \implies \mathscr{E}_4 \subset \mathscr{M}(\mathscr{E}_6)$$
$$[a,b) = [a,\infty) \cap [b,\infty)^c \in \mathscr{M}(\mathscr{E}_7) \implies \mathscr{E}_6 \subset \mathscr{M}(\mathscr{E}_7)$$
$$(a,b] = (-\infty,a]^c \cap (-\infty,b] \in \mathscr{M}(\mathscr{E}_8) \implies \mathscr{E}_3 \subset \mathscr{M}(\mathscr{E}_8)$$

□

Product σ-Algebra

Let $\{X_i\}_{i\in I}$ be non-empty sets, $X = \prod_{i\in I} X_i$, and $\pi_i \colon X \to X_i$ the coordinate maps. Let $\mathscr{M}_i$ be the σ-algebra on X_i for each i.

Definition 28.8 (Product σ-algebra). The ***product σ-algebra*** on X, denoted by $\bigotimes_{i\in I} \mathscr{M}_i$, is the σ-algebra generated by

$$\left\{\pi_i^{-1}(E_i) \mid E_i \in \mathscr{M}_i,\, i \in I\right\}.$$

If $I = \{1,\dots,n\}$, we also write

$$\bigotimes_{i=1}^{n} \mathscr{M}_i \quad \text{or} \quad \mathscr{M}_1 \otimes \cdots \otimes \mathscr{M}_n.$$

Proposition 28.9. *If I is countable, then $\bigotimes_{i\in I} \mathscr{M}_i$ is the σ-algebra generated by $\{\prod_{i\in I} E_i \mid E_i \in \mathscr{M}_i\}$.*

Proof. If $E_i \in \mathscr{M}_i$, then $\pi_i^{-1}(E_i) = \prod_{\beta\in A} E_\beta$ where $E_\beta = X$ for $\beta \neq \alpha$; on the other hand, $\prod_{i\in I} E_i = \bigcap_{i\in I} \pi_i^{-1}(E_i)$. The result therefore follows from 28.5. □

Proposition 28.10. *Suppose $\mathscr{M}_i = \mathscr{M}(\mathscr{E}_i)$ for each $i \in I$.*

(i) $\bigotimes_{i\in I} \mathscr{M}_i$ is generated by $\mathscr{F}_1 = \{\pi_i^{-1}(E_i) \mid E_i \in \mathscr{E}_i,\, i \in I\}$.

(ii) If I is countable and $X_i \in \mathscr{E}_i$ for all i, then $\bigotimes_{i\in I} \mathscr{M}_i$ is generated by $\mathscr{F}_2 = \{\prod_{i\in I} E_i \mid$

$E_i \in \mathscr{E}_i\}$.

Proof.

(i) Since $\mathscr{F}_1 \subset \left\{\pi_i^{-1}(E_i) \mid E_i \in \mathscr{M}_i,\, i \in I\right\}$, by 28.5, we have $\mathscr{M}(\mathscr{F}_1) \subset \bigotimes_{i\in I} \mathscr{M}_i$.

We now show the reverse inclusion. For each $i \in I$, let

$$\mathscr{N}_i = \{E \subset X_i \mid \pi_i^{-1}(E) \in \mathscr{M}(\mathscr{F}_1)\}.$$

We check that $\mathscr{N}_i$ is a σ-algebra on X_i.

(i) Let $E_1, E_2, \cdots \in \mathscr{N}_i$. Then

$$\pi_i^{-1}\left(\bigcup_{n=1}^{\infty} E_n\right) = \bigcup_{n=1}^{\infty} \pi_i^{-1}(E_n) \in \mathscr{M}(\mathscr{F}_1) \implies \bigcup_{n=1}^{\infty} E_n \in \mathscr{N}_i.$$

(ii) Let $E \in \mathscr{N}_i$. Then

$$\pi_i^{-1}(E^c) = [\pi_i^{-1}(E)]^c \in \mathscr{M}(\mathscr{F}_1) \implies E^c \in \mathscr{N}_i.$$

Thus $\mathscr{N}_i$ contains $\mathscr{E}_i$, and hence $\mathscr{M}_i$. In other words, $\pi_i^{-1}(E) \in \mathscr{M}(\mathscr{F}_1)$ for all $E \in \mathscr{M}_i$, and hence $\bigotimes_{i\in I} \mathscr{M}_i \subset \mathscr{M}(\mathscr{F}_1)$.

(ii) The second assertion follows from the first as in the proof of Proposition 1.3.

□

Proposition 28.11. *Let $X_1, \ldots, X_n$ be metric spaces and let $X = \prod_{i=1}^n X_i$, equipped with the product metric. Then $\bigotimes_{i=1}^n \mathscr{B}(X_i) \subset \mathscr{B}(X)$. If the X_i's are separable, then $\bigotimes_{i=1}^n \mathscr{B}(X_i) = \mathscr{B}(X)$.*

Proof. By Proposition 1.4, $\bigotimes_{i=1}^n \mathscr{B}(X_i)$ is generated by the sets $\pi_i^{-1}(U_i)$, $1 \le i \le n$, where U_i is open in X_i. Since these sets are open in X, 28.5 implies that $\bigotimes_{i=1}^n \mathscr{B}(X_i) \subset \mathscr{B}(X)$.

Suppose now that C_i is a countable dense set in X_i, and let $\mathscr{E}_i$ be the collection of balls in X_i with rational radius and center in C_i. Then every open set in X_i is a union of members of $\mathscr{E}_i$ — in fact, a countable union since $\mathscr{E}_i$ is itself countable. Moreover, the set of points in X whose i-th coordinate is in C_i for all i is a countable dense subset of X, and the balls of radius r in X are merely products of balls of radius r in the X_i's.

It follows that $\mathscr{B}(X_i)$ is generated by $\mathscr{E}_i$ and $\mathscr{B}(X)$ is generated by $\{\prod_{i=1}^n E_i \mid E_i \in \mathscr{E}_i\}$.

Therefore $\mathscr{B}(X) = \bigotimes_{i=1}^n \mathscr{B}(X_i)$ by Proposition 1.4. □

Corollary 28.12. $\mathscr{B}(\mathbb{R}^n) = \bigotimes_{i=1}^n \mathscr{B}(\mathbb{R})$.

We conclude this section with a technical result that will be needed later.

Definition 28.13. An ***elementary family*** is a collection $\mathscr{E} \subset \mathscr{P}(X)$ that satisfies

(i) $\emptyset \in \mathscr{E}$;

(ii) if $E, F \in \mathscr{E}$ then $E \cap F \in \mathscr{E}$;

(iii) if $E \in \mathscr{E}$ then E^c is a finite disjoint union of members of $\mathscr{E}$.

Proposition 28.14. *If $\mathscr{E}$ is an elementary family, the collection $\mathscr{A}$ of finite disjoint unions of members of $\mathscr{E}$ is an algebra.*

Proof. We check that $\mathscr{A}$ is an algebra:

(i)

(ii) Let $A \in \mathscr{A}$. Then $A = \bigcup_{i=1}^n A_i$ where $A_1, \ldots, A_n \in \mathscr{E}$. For each $i = 1, \ldots, n$, let ${A_i}^c = \bigcup_{j=1}^{J_i} B_{i,j}$ where $B_{i,1}, \ldots, B_{i,J_i}$ are disjoint members of $\mathscr{E}$. Then

$$\left(\bigcup_{i=1}^n A_i \right)^c = \bigcap_{i=1}^n {A_i}^c = \bigcap_{i=1}^n \left(\bigcup_{j=1}^{J_i} B_{i,j} \right) =$$

(iii)

□

28.3 Measures

Definition 28.15 (Measure). Let X be a set equipped with a σ-algebra $\mathscr{M}$. A ***measure*** on X is a function $\mu : \mathscr{M} \to [0,\infty]$ such that

(i) $\mu(\emptyset) = 0$;

(ii) $\mu\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \mu(E_n)$ for every disjoint $E_1, E_2, \cdots \in \mathscr{M}$.(countable additivity)

We call $(X, \mathscr{M}, \mu)$ a ***measure space***.

Countable additivity implies *finite additivity*:

(ii') if $E_1, \ldots, E_n$ are disjoint sets in $\mathscr{M}$, then $\mu(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} \mu(E_i)$,

because one can take $E_{n+1} = E_{n+2} = \cdots = \emptyset$. If μ satisfies (i) and (ii') but not necessarily (ii), we call μ a *finitely additive measure*.

We introduce some standard terminology concerning the "size" of μ.

- If $\mu(X) < \infty$ (which implies that $\mu(E) < \infty$ for all $E \in \mathscr{M}$), we say μ is *finite*.
- If $X = \bigcup_{n=1}^{\infty} E_n$ where $E_n \in \mathscr{M}$ and $\mu(E_n) < \infty$ for all n, we say μ is σ*-finite*. (That is, the whole space can be covered by measurable sets of finite measure.)
- More generally, if $E = \bigcup_{n=1}^{\infty} E_n$ where $E_n \in \mathscr{M}$ and $\mu(E_n) < \infty$ for all i, we say the set E is σ*-finite* for μ.
- If for each $E \in \mathscr{M}$ with $\mu(E) = \infty$ there exists $F \in \mathscr{M}$ with $F \subset E$ and $0 < \mu(F) < \infty$, μ is called *semifinite*.

Example.

- For any $E \subset X$, where X is any set, define $\mu(E) = \infty$ if E is an infinite set, and let $\mu(E)$ be the number of points in E if E is finite. This μ is called the *counting measure* on X.
- Fix $x_0 \in X$. For any $E \subset X$, let

$$\delta_{x_0}(E) = \begin{cases} 1 & (x_0 \in E) \\ 0 & (x_0 \notin E) \end{cases}$$

  This measure is called the *unit mass* concentrated at x_0 (or the *Dirac measure*).
- A *probability measure* on Ω is a measure $\mathbb{P}$ such that $\mathbb{P}(\Omega) = 1$.

The basic properties of measures are summarised in the following result.

Lemma 28.16. *Suppose $(X, \mathscr{M}, \mu)$ is a measure space.*

(i) If $E, F \in \mathscr{M}$ and $E \subset F$, then $\mu(E) \leq \mu(F)$. *(monotonicity)*

(ii) If $E_1, E_2, \cdots \in \mathscr{M}$, then $\mu\left(\bigcup_{n=1}^{\infty} E_i\right) \leq \sum_{n=1}^{\infty} \mu(E_n)$. *(subadditivity)*

(iii) If $E_1, E_2, \cdots \in \mathscr{M}$ and $E_1 \subset E_2 \subset \cdots$, then *(continuity from below)*

$$\mu\left(\bigcup_{n=1}^{\infty} E_n\right) = \lim_{n\to\infty} \mu(E_n).$$

(iv) If $E_1, E_2, \cdots \in \mathscr{M}$, $E_1 \supset E_2 \supset \cdots$, and $\mu(E_1) < \infty$, then *(continuity from above)*

$$\mu\left(\bigcap_{n=1}^{\infty} E_n\right) = \lim_{n\to\infty} \mu(E_n).$$

Proof.

(i) If $E \subset F$, note that $F = E \cup (F \setminus E)$. Then

$$\mu(F) = \mu(E) + \mu(F \setminus E) \geq \mu(E).$$

(ii) Let $F_1 = E_1$ and $F_n = E_n \setminus \left(\bigcup_{i=1}^{n-1} E_i\right)$ for $n > 1$. Then the F_n's are disjoint and $\bigcup_{i=1}^{n} F_i = \bigcup_{i=1}^{n} E_i$ for all n. Hence by (i),

$$\begin{aligned}
\mu\left(\bigcup_{n=1}^{\infty} E_n\right) &= \mu\left(\bigcup_{n=1}^{\infty} F_n\right) \\
&= \sum_{n=1}^{\infty} \mu(F_n) && \text{[by countable additivity]} \\
&\leq \sum_{n=1}^{\infty} \mu(E_n) && \text{[by monotonicity, since } F_n \subset E_n\text{]}
\end{aligned}$$

(iii) Suppose $E_1 \subset E_2 \subset \cdots$. Then we have

$$\begin{aligned}
\mu\left(\bigcup_{n=1}^{\infty} E_n\right) &= \mu\left(\bigcup_{n=1}^{\infty} E_n \setminus E_{n-1}\right) && \text{[setting } E_0 = \emptyset\text{]} \\
&= \sum_{n=1}^{\infty} \mu(E_n \setminus E_{n-1}) && \text{[by countable additivity]} \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} \mu(E_i \setminus E_{i-1}) \\
&= \lim_{n\to\infty} \mu(E_n) && \text{[by finite additivity]}
\end{aligned}$$

(iv) Let $F_n = E_1 \setminus E_n$. Then $F_1 \subset F_2 \subset \cdots$. Note that $\mu(E_1) = \mu(F_n) + \mu(E_n)$, and

$$\bigcup_{n=1}^{\infty} F_n = E_1 \setminus \left(\bigcap_{n=1}^{\infty} E_n\right).$$

Take the measure on both sides of the equation. By (iii),

$$\mu\left(\bigcup_{n=1}^{\infty} F_n\right) = \lim_{n\to\infty} \mu(F_n),$$

so

$$\begin{aligned}
\mu(E_1) &= \mu\left(\bigcap_{n=1}^{\infty} E_n\right) + \lim_{n\to\infty} \mu(F_n) \\
&= \mu\left(\bigcap_{n=1}^{\infty} E_n\right) + \lim_{n\to\infty} [\mu(E_1) - \mu(E_n)] \\
&= \mu\left(\bigcap_{n=1}^{\infty} E_n\right) + \mu(E_1) - \lim_{n\to\infty} \mu(E_n)
\end{aligned}$$

Since $\mu(E_1) < \infty$, we may subtract it from both sides to yield the desired result.

□

In (iv), the condition that $\mu(E_1) < \infty$ cannot be omitted. **Counterexample:** $E_n = [n, \infty)$.

The next result is often known as the *inclusion–exclusion formula*, which allows us to compute the measure of a union.

Lemma 28.17. *Suppose $(X, \mathscr{M}, \mu)$ is a measure space. If $E, F \in \mathscr{M}$, then*

$$\mu(E \cup F) = \mu(E) + \mu(F) - \mu(E \cap F).$$

Proof. We have

$$E \cup F = (E \setminus (E \cap F)) \cup (F \setminus (E \cap F)) \cup (E \cap F)$$

where the RHS is a disjoint union. Hence

$$\begin{aligned} \mu(E \cup F) &= \mu(E \setminus (E \cap F)) + \mu(F \setminus (E \cap F)) + \mu(E \cap F) \\ &= (\mu(E) - \mu(E \cap F)) + (\mu(F) - \mu(E \cap F)) + \mu(E \cap F) \\ &= \mu(E) + \mu(F) - \mu(E \cap F). \end{aligned}$$

□

Suppose $(X, \mathscr{M}, \mu)$ is a measure space. We say $E \in \mathscr{M}$ is a ***null set*** if $\mu(E) = 0$. By subadditivity, any countable union of null sets is a null set, a fact which we shall use frequently.

If a statement about points $x \in X$ is true except for x in some null set, we say that it is true ***almost everywhere*** (a.e.). (If more precision is needed, we shall speak of a μ-null set, or μ-almost everywhere.)

If $\mu(E) = 0$ and $F \subset E$, then $\mu(F) = 0$ by monotonicity provided that $F \in \mathscr{M}$, but in general it need not be true that $F \in \mathscr{M}$. In the case when this is true, we make the following definition:

Definition 28.18. We say μ is a ***complete measure*** on $\mathscr{M}$ if $\mathscr{M}$ contains all subsets of null sets:

$$F \subset E \in \mathscr{M} \text{ and } \mu(E) = 0 \implies F \in \mathscr{M}.$$

The next result states that completeness can be achieved by enlarging the domain of μ.

Theorem 28.19. *Suppose $(X, \mathscr{M}, \mu)$ is a measure space. Let $\mathscr{N} = \{N \in \mathscr{M} \mid \mu(N) = 0\}$ and*

$$\overline{\mathscr{M}} = \{E \cup F \mid E \in \mathscr{M}, F \subset N \text{ for some } N \in \mathscr{N}\}.$$

Then $\overline{\mathscr{M}}$ is a σ-algebra, and there exist a unique extension $\overline{\mu}$ of μ to a complete measure on $\overline{\mathscr{M}}$.

We call $\overline{\mu}$ the *completion* of μ, and call $\overline{\mathscr{M}}$ the *completion* of $\mathscr{M}$ with respect to μ.

Proof. We check that $\overline{\mathscr{M}}$ is a σ-algebra.

(i) Let $E \cup F \in \overline{\mathscr{M}}$, where $E \in \mathscr{M}$ and $F \subset N \in \mathscr{N}$. We can assume that $E \cap N = \emptyset$ (otherwise replace F and N by $F \setminus E$ and $N \setminus E$). Then $E \cup F = (E \cup N) \cap (N^c \cup F)$, so $(E \cup F)^c = (E \cup N)^c \cup (N \setminus F)$. But $(E \cup N)^c \in \mathscr{M}$ and $N \setminus F \subset N$, so that $(E \cup F)^c \in \overline{\mathscr{M}}$.

(ii) Since $\mathscr{M}$ and $\mathscr{N}$ are closed under countable unions, this implies $\overline{\mathscr{M}}$ is closed under countable unions.

Define $\overline{\mu}\colon \overline{\mathscr{M}} \to [0,\infty]$ by

$$\overline{\mu}(E \cup F) = \mu(E)$$

for each $E \cup F \in \overline{\mathscr{M}}$. This is well-defined, since if $E_1 \cup F_1 = E_2 \cup F_2$ where $F_i \subset N_i \in \mathscr{N}$, then $E_1 \subset E_2 \cup N_2$ and so $\mu(E_1) \leq \mu(E_2) + \mu(N_2) = \mu(E_2)$, and likewise $\mu(E_2) \leq \mu(E_1)$; thus $\mu(E_1) = \mu(E_2)$.

It is easily verified that $\overline{\mu}$ is a complete measure on $\overline{\mathscr{M}}$, and that $\overline{\mu}$ is the only measure on $\overline{\mathscr{M}}$ that extends μ. □

28.4 Outer Measures

In this section we develop the tools we shall use to construct measures.

To motivate the ideas, recall the procedure used in calculus to define the area of a bounded region $E \subset \mathbb{R}^2$: draw a grid of rectangles in the plane, and approximate the area of E from below and above by the sum of areas of rectangles. Then take the limits of these approximations by making the grid finer and finer, giving the "inner area" and "outer area" of E; if they are equal, the common value is the "area" of E.

The abstract generalisation of the notion of outer area is as follows:

Definition 28.20 (Outer measure)**.** An ***outer measure*** on a non-empty set X is a function $\mu^*\colon \mathscr{P}(X) \to [0,\infty]$ that satisfies

(i) $\mu^*(\emptyset) = 0$;

(ii) $\mu^*(A) \leq \mu^*(B)$ if $A \subset B$; (monotonicity)

(iii) $\mu^*\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mu^*(A_n)$. (countable subadditivity)

The most common way to obtain outer measures is to start with a family $\mathscr{E}$ of "elementary sets" on which a notion of measure is defined (such as rectangles in the plane) and then to approximate arbitrary sets "from the outside" by countable unions of members of $\mathscr{E}$. The precise construction is as follows.

Proposition 28.21. *Let $\mathscr{E} \subset \mathscr{P}(X)$ be such that $\emptyset \in \mathscr{E}$, $X \in \mathscr{E}$. Let $\rho\colon \mathscr{E} \to [0,\infty]$ be such that $\rho(\emptyset) = 0$. Define $\mu^*\colon \mathscr{P}(X) \to [0,\infty]$ by*

$$\mu^*(A) := \inf\left\{ \sum_{n=1}^{\infty} \rho(E_n) \,\middle|\, E_n \in \mathscr{E} \text{ and } A \subset \bigcup_{n=1}^{\infty} E_n \right\}. \tag{28.1}$$

Then μ^* is an outer measure.

Proof. For any $A \subset X$, there exists (E_n) in $\mathscr{E}$ such that $A \subset \bigcup_{n=1}^{\infty} E_n$ (take $E_n = X$ for all n), so the definition of μ^* makes sense.

Now we check that μ^* is an outer measure:

(i) We have $\emptyset \subset \bigcup_{n=1}^{\infty} \emptyset$, so $0 \leq \mu^*(\emptyset) \leq \sum_{n=1}^{\infty} \rho(\emptyset) = 0$ implies $\mu^*(\emptyset) = 0$.

(ii) Let $A, B \in \mathscr{P}(X)$ be such that $A \subset B$. There exist $E_1, E_2, \cdots \in \mathscr{E}$ such that $B \subset \bigcup_{n=1}^{\infty} E_n$. Thus

$$A \subset B \subset \bigcup_{n=1}^{\infty} E_n \implies \mu^*(A) \leq \sum_{n=1}^{\infty} \rho(E_n) \implies \mu^*(A) \leq \mu^*(B)$$

by taking infimum.

(iii) Let $A_1, A_2, \ldots, \in \mathscr{P}(X)$. Let $\varepsilon > 0$ be given. For each n, there exist $E_{n,1}, E_{n,2}, \cdots \in \mathscr{E}$ such that $A_n \subset \bigcup_{k=1}^{\infty} E_{n,k}$ and

$$\mu^*(A_n) \leq \sum_{k=1}^{\infty} \rho(E_{n,k}) \leq \mu^*(A_n) + \frac{\varepsilon}{2^n}.$$

Then

$$A := \bigcup_{n=1}^{\infty} A_n \subset \bigcup_{n=1}^{\infty} \bigcup_{k=1}^{\infty} E_{n,k} \implies \mu^*(A) \leq \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \rho(E_{n,k}) \leq \sum_{n=1}^{\infty} \mu^*(A_n) + \frac{\varepsilon}{2^n} = \sum_{n=1}^{\infty} \mu^*(A_n) + \varepsilon.$$

Since ε is arbitrary, we conclude that $\mu^*(A) \leq \sum_{n=1}^{\infty} \mu^*(A_n)$.

□

The fundamental step that leads from outer measures to measures is as follows.

Definition 28.22. Let μ^* be an outer measure on X. We say $A \subset X$ is ***μ^*-measurable*** if

$$\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^c)$$

for all $E \subset X$.

By subadditivity, the inequality $\mu^*(E) \leq \mu^*(E \cap A) + \mu^*(E \cap A^c)$ holds for any A and E, so to prove that A is μ^*-measurable, it suffices to prove the reverse inequality. The latter is trivial if $\mu^*(E) = \infty$, so we see that A is μ^*-measurable iff

$$\mu^*(E) \geq \mu^*(E \cap A) + \mu^*(E \cap A^c)$$

for all $E \subset X$ such that $\mu^*(E) < \infty$.

Some motivation for the notion of μ^*-measurability can be obtained by referring to the discussion at the beginning of this section. If E is a "well-behaved" set such that $E \supset A$, the equation $\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^c)$ says that the outer measure of A, $\mu^*(A)$ is equal to the "inner measure" of A, $\mu^*(E \cap A) + \mu^*(E \cap A^c)$. The leap from "well-behaved" sets containing A to arbitrary subsets of X a large one, but it is justified by the following theorem.

Theorem 28.23 (Carathéodory's extension theorem). *If μ^* is an outer measure on X, the collection $\mathscr{M}$ of μ^*-measurable sets is a σ-algebra, and $\mu^*|_{\mathscr{M}}$ is a complete measure.*

Proof. □

Our first applications of Carathéodory's extension theorem will be in the context of extending measures from algebras to σ-algebras.

Definition 28.24 (Premeasure)**.** Let $\mathscr{A} \subset \mathscr{P}(X)$ be an algebra. A ***premeasure*** is a function $\mu_0 \colon \mathscr{A} \to [0,\infty]$ that satisfies

(i) $\mu_0(\emptyset) = 0$;

(ii) $\mu_0\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu_0(A_n)$ for every $A_1, A_2, \cdots \in \mathscr{A}$ such that $\bigcup_{n=1}^{\infty} A_n \in \mathscr{A}$.

Proposition 28.25. *Let μ_0 be a premeasure on $\mathscr{A}$, and let μ^* be defined as in (28.1). Then*

(i) $\mu^*|_{\mathscr{A}} = \mu_0$;

(ii) every set in $\mathscr{A}$ is μ^*-measurable.

Theorem 28.26. *Let $\mathscr{A} \subset \mathscr{P}(X)$ be an algebra, μ_0 a premeasure on $\mathscr{A}$, and $\mathscr{M}$ the σ-algebra generated by $\mathscr{A}$. There exists a measure μ on $\mathscr{M}$ whose restriction to $\mathscr{A}$ is μ_0 – namely, $\mu = \mu^*|_{\mathscr{M}}$ where μ^* is defined above.*
If ν is another measure on $\mathscr{M}$ that extends μ_0, then $\nu(E) \leq \mu(E)$ for all $E \in \mathscr{M}$, with equality when $\mu(E) < \infty$.

If μ_0 is σ-finite, then μ is the unique extension of μ_0 to a measure on $\mathscr{M}$.

28.5 Borel Measure on $\mathbb{R}$

We are now in a position to construct a definitive theory for measuring subsets of $\mathbb{R}$ based on the idea that the measure of an interval is its length. We begin with a more general construction that yields a large family of measures on $\mathbb{R}$ whose domain is the Borel σ-algebra $\mathscr{B}(\mathbb{R})$; such measures are called *Borel measures* on $\mathbb{R}$.

Lebesgue–Stieltjes Measure

Definition 28.27 (Borel measure)**.** A ***Borel measure*** is a measure whose domain is a Borel σ-algebra.

Suppose $\mu : \mathscr{B}(\mathbb{R}) \to [0,\infty]$ is a finite Borel measure on $\mathbb{R}$. Define

$$F(x) := \mu([-\infty, x]).$$

We call F the *distribution function* of μ.

Proposition 28.28. *Let $F : \mathbb{R} \to \mathbb{R}$ be increasing and right continuous. If $(a_i, b_i]$ are disjoint h-intervals, define*

$$\mu_0\left(\bigcup_{i=1}^{n}(a_i,b_i]\right) := \sum_{i=1}^{n}[F(b_i) - F(a_i)],$$

and let $\mu_0(\emptyset) = 0$. Then μ_0 is a premeasure on the algebra $\mathscr{A}$.

Theorem 28.29.

Definition 28.30 (Lebesgue–Stieltjes measure)**.** We call μ_F the ***Lebesgue–Stieltjes measure*** associated to F.

Lebesgue-Stieltjes measures enjoy some useful regularity properties that we now investigate. In this discussion, we fix a complete Lebesgue-Stieltjes measure μ on $\mathbb{R}$ associated to the increasing, right continuous function F, and we denote by $\mathscr{M}_\mu$ the domain of μ. Thus, for any $E \in \mathscr{M}_\mu$,

$$\begin{aligned}\mu(E) &= \inf\left\{\sum_{n=1}^{\infty}[F(b_n) - F(a_n)] \,\middle|\, E \subset \bigcup_{n=1}^{\infty}(a_n,b_n]\right\} \\ &= \inf\left\{\sum_{n=1}^{\infty}\mu((a_n,b_n]) \,\middle|\, E \subset \bigcup_{n=1}^{\infty}(a_n,b_n]\right\}.\end{aligned}$$

We first observe that in the second formula for $\mu(E)$, we can replace h-intervals by open h-intervals:

Lemma 28.31. *For any $E \in \mathscr{M}_\mu$,*

$$\mu(E) = \inf\left\{ \sum_{n=1}^{\infty} \mu((a_n, b_n)) \,\middle|\, E \subset \bigcup_{n=1}^{\infty} (a_i, b_i) \right\}.$$

Theorem 28.32. *If $E \in \mathscr{M}_\mu$, then*

$$\mu(E) = \inf_{\substack{U \supset E, \\ U \text{ is open}}} \mu(U) = \sup_{\substack{K \subset E, \\ K \text{ is compact}}} \mu(K).$$

Theorem 28.33. *If $E \subset \mathbb{R}$, the following are equivalent:*

(i) $E \in \mathscr{M}_\mu$.

(ii) $E = V \setminus N_1$, *where V is a G_δ set and $\mu(N_1) = 0$.*

(iii) $E = H \cup N_2$, *where H is an F_σ set and $\mu(N_2) = 0$.*

Proposition 28.34.

Lebesgue Measure

We now examine the most important measure on $\mathbb{R}$, namely, Lebesgue measure:

Definition 28.35 (Lebesgue measure)**.**

Among the most significant properties of Lebesgue measure are its invariance under translations and simple behaviour under dilations.

If $E \subset \mathbb{R}$ and $s, r \in \mathbb{R}$, we define

$$E + s := \{x + s \mid x \in E\}, \quad rE = \{rx \mid x \in E\}.$$

Theorem 28.36. *Let $E \in \mathscr{L}$, and $s, r \in \mathbb{R}$. Then $E + s \in \mathscr{L}$ and $rE \in \mathscr{L}$. Moreover,*

(i) $m(E + s) = m(E)$; *(translation invariant)*

(ii) $m(rE) = |r| m(E)$.

Cantor Set and Cantor Function

Definition 28.37 (Cantor set).

Thus C is obtained from $[0,1]$ by removing the open middle third $\left(\frac{1}{3},\frac{2}{3}\right)$, then removing the open middle thirds $\left(\frac{1}{9},\frac{2}{9}\right)$ and $\left(\frac{7}{9},\frac{8}{9}\right)$ of the two remaining intervals, and so forth.

The basic properties of C are summarised as follows:

Proposition 28.38. *Let C be the Cantor set.*

(i) *C is compact, nowhere dense, and totally disconnected (i.e., the only connected subsets of C are single points). Moreover, C has no isolated points.*

(ii) $m(C) = 0$.

(iii) *C is uncountable.*

Proof.

(i)

(ii) By continuity from above,

$$m(C) = \lim_{n\to\infty} m(C_n) = \lim_{n\to\infty} \left(\frac{2}{3}\right)^n = 0.$$

(iii)

□

Definition 28.39 (Cantor function).

28.6 Lebesgue Measure

Additivity of Outer Measure on Borel Sets

Recall that there exist disjoint $E, F \subset \mathbb{R}$ such that $|E \cup F| \neq |E| + |F|$. Thus outer measure, despite its name, is not a measure on the σ-algebra of all subsets of $\mathbb{R}$.

Our main goal in this section is to prove that outer measure, when restricted to the Borel subsets of $\mathbb{R}$, is a measure.

The next result is our first step toward the goal of proving that outer measure restricted to the Borel sets is a measure. It states that the outer measure is additive if one of the sets is open.

Lemma 28.40. *Suppose $E, G \subset \mathbb{R}$ are disjoint, and G is open. Then*

$$|E \cup G| = |E| + |G|.$$

Proof. If $|G| = \infty$, then $|E \cup G| = \infty$ and $|E| + |G| = \infty$. Thus assume $|G| < \infty$.

By subadditivity, $|E \cup G| \leq |E| + |G|$. Thus we need to prove the inequality only in the other direction.

Lemma. *If G is a union of m disjoint open intervals that are all disjoint from E, then $|E \cup G| = |E| + |G|$.*

Proof. We first prove the case $m = 1$. Then $G = (a, b)$ for some $a, b \in \mathbb{R}$ with $a < b$.

We can assume $a, b \notin E$, since this does not change its outer measure.

Let (I_n) be a sequence of open intervals whose union contains $E \cup G$. For each n, let

$$J_n = I_n \cap (-\infty, a), \quad K_n = I_n \cap (a, b), \quad L_n = I_n \cap (b, \infty).$$

Then

$$\ell(I_n) = \ell(J_n) + \ell(K_n) + \ell(L_n).$$

Now $J_1, L_1, J_2, L_2, \ldots$ is a sequence of open intervals whose union contains E, and $K_1, K_2, \ldots$ is a sequence of open intervals whose union contains G. Thus

$$\begin{aligned} \sum_{n=1}^{\infty} \ell(I_n) &= \sum_{n=1}^{\infty} (\ell(J_n) + \ell(L_n)) + \sum_{n=1}^{\infty} \ell(K_n) \\ &\geq |E| + |G|. \end{aligned}$$

The inequality above implies that $|E \cup G| \geq |E| + |G|$. Hence $|E \cup G| = |E| + |G|$.

Then induct on m to prove the lemma. □

Suppose G is an arbitrary open subset of $\mathbb{R}$ disjoint from E. Then $G = \bigcup_{n=1}^{\infty} I_n$ for some sequence of disjoint intervals $I_1, I_2, \ldots$, each of which is disjoint from E.

For each m, by the lemma,

$$|E \cup G| \geq \left| E \cup \left(\bigcup_{n=1}^{m} I_n \right) \right| = |E| + \left| \bigcup_{n=1}^{m} I_n \right| = |E| + \sum_{n=1}^{m} \ell(I_n).$$

Since this holds for all m, we have

$$|E \cup G| \geq |E| + \sum_{n=1}^{\infty} \ell(I_n) \geq |E| + |G|.$$

□

The next result shows that the outer measure is additive if at least one of the sets is closed.

Lemma 28.41. *Suppose $E, F \subset \mathbb{R}$ are disjoint, and F is closed. Then*

$$|E \cup F| = |E| + |F|.$$

Proof. By subadditivity, $|E \cup F| \leq |E| + |F|$. Thus it suffices to prove $|E| + |F| \leq |E \cup F|$.

□

Recall that the collection of Borel sets is the smallest σ-algebra on $\mathbb{R}$ that contains all open subsets of $\mathbb{R}$. The next result provides an extremely useful tool for approximating a Borel set by a closed set.

Lemma 28.42. *Suppose $B \subset \mathbb{R}$ is a Borel set. Then*

$$\forall \varepsilon > 0, \quad \exists \text{ closed } F \subset B, \quad |B \setminus F| < \varepsilon.$$

Now we can prove that the outer measure of the disjoint union of two sets is what we expect if at least one of the two sets is a Borel set.

Lemma 28.43. *Suppose $E, B \subset \mathbb{R}$ are disjoint, and B is a Borel set. Then*

$$|E \cup B| = |E| + |B|.$$

You have probably long suspected that not every subset of $\mathbb{R}$ is a Borel set. Now we can prove this suspicion.

Corollary 28.44. *There exists $B \subset \mathbb{R}$ such that $|B| < \infty$ and B is not a Borel set.*

The tools we have constructed now allow us to prove that outer measure, when restricted to the Borel sets, is a measure.

Proposition 28.45. *Outer measure is a measure on* $(\mathbb{R}, \mathscr{B})$.

The result above implies that the next definition makes sense.

Definition 28.46 (Lebesgue measure)**.** ***Lebesgue measure*** is the measure on $(\mathbb{R}, \mathscr{B})$ that assigns to each Borel set its outer measure.

Lebesgue Measurable Sets

Cantor Set and Cantor Function

29 Integration

29.1 Measurable Functions

Let $(X, \mathscr{M})$ and $(Y, \mathscr{N})$ be measurable spaces.

Definition 29.1 (Measurable function)**.** We say $f\colon X \to Y$ is $(\mathscr{M}, \mathscr{N})$***-measurable*** if $f^{-1}(E) \in \mathscr{M}$ for every $E \in \mathscr{N}$.

In other words, f is measurable if the pre-image of every measurable set is a measurable set.

Notation. If $\mathscr{M}$ and $\mathscr{N}$ are understood, we simply write $f\colon X \to Y$ is measurable.

Lemma 29.2. *The composition of measurable functions is measurable.*

Proof. Suppose $f\colon X \to Y$ and $g\colon Y \to Z$ are measurable functions. Let $E \subset Z$ be measurable. Then

$$(g \circ f)^{-1}(E) = f^{-1}(g^{-1}(E)).$$

Since g is measurable, $g^{-1}(E)$ is measurable. Since f is measurable, $f^{-1}(g^{-1}(E))$ is measurable. □

Proposition 29.3. *If $\mathscr{N}$ is generated by $\mathscr{E}$, then $f\colon X \to Y$ is $(\mathscr{M}, \mathscr{N})$-measurable if and only if $f^{-1}(E) \in \mathscr{M}$ for all $E \in \mathscr{E}$.*

Proof.

$\boxed{\Longrightarrow}$ Trivial.

$\boxed{\Longleftarrow}$ Suppose $f^{-1}(E) \in \mathscr{M}$ for all $E \in \mathscr{E}$.

Claim. $\mathscr{S} = \{E \subset Y \mid f^{-1}(E) \in \mathscr{M}\}$ is a σ-algebra on Y.

(i) $\emptyset \in \mathscr{S}$, because $f^{-1}(\emptyset) = \emptyset \in \mathscr{M}$.

$Y \in \mathscr{S}$, because $f^{-1}(Y) = X \in \mathscr{M}$.

(ii) Let $E \in \mathscr{S}$. Then $f^{-1}(E) \in \mathscr{M}$, so $[f^{-1}(E)]^c = f^{-1}(E^c) \in \mathscr{M}$. Thus $E^c \in \mathscr{S}$.

(iii) Let $E_1, E_2, \cdots \in \mathscr{S}$. Then $f^{-1}(E_1), f^{-1}(E_2), \cdots \in \mathscr{M}$, so $\bigcup_{n=1}^{\infty} f^{-1}(E_n) = f^{-1}\left(\bigcup_{n=1}^{\infty} E_n\right) \in \mathscr{M}$. Thus $\bigcup_{n=1}^{\infty} E_n \in \mathscr{S}$.

Since $\mathscr{E} \subset \mathscr{S}$, by 28.5, this implies $\mathscr{N} = \mathscr{M}(\mathscr{E}) \subset \mathscr{S}$. □

Corollary 29.4. *If X and Y are metric (or topological) spaces, every continuous $f: X \to Y$ is $(\mathscr{B}(X), \mathscr{B}(Y))$-measurable.*

Proof. f is continuous iff $f^{-1}(U)$ is open in X for every open $U \subset Y$. □

Definition 29.5. Let $(X, \mathscr{M})$ be a measurable space. A real- or complex-valued function f on X is called ***$\mathscr{M}$-measurable***, or just measurable, if it is $(\mathscr{M}, \mathscr{B}(\mathbb{R}))$- or $(\mathscr{M}, \mathscr{B}(\mathbb{C}))$-measurable.
In particular, $f: \mathbb{R} \to \mathbb{C}$ is ***Lebesgue*** (resp. ***Borel***) ***measurable*** if it is $(\mathscr{L}, \mathscr{B}(\mathbb{C}))$ (resp. $(\mathscr{B}(\mathbb{R}), \mathscr{B}(\mathbb{C}))$) measurable; likewise for $f: \mathbb{R} \to \mathbb{R}$.

Proposition 29.6. *Let $(X, \mathscr{M})$ be a measurable space, and $f: X \to \mathbb{R}$. Then the following are equivalent:*

(i) f is $\mathscr{M}$-measurable.

(ii) $f^{-1}((a, \infty)) \in \mathscr{M}$ for all $a \in \mathbb{R}$.

(iii) $f^{-1}([a, \infty)) \in \mathscr{M}$ for all $a \in \mathbb{R}$.

(iv) $f^{-1}((-\infty, a)) \in \mathscr{M}$ for all $a \in \mathbb{R}$.

(v) $f^{-1}((-\infty, a]) \in \mathscr{M}$ for all $a \in \mathbb{R}$.

Proof.

(i) $\Longrightarrow$ (ii), (iii), (iv), (v) The sets (a, ∞), $[a, \infty)$, $(-\infty, a)$, and $(-\infty, a]$ are all Borel sets.

(ii), (iii), (iv), (v) $\Longrightarrow$ (i) By 28.7, the sets (a, ∞), $[a, \infty)$, $(-\infty, a)$, and $(-\infty, a]$ generate $\mathscr{B}(\mathbb{R})$. Hence by 29.3, we conclude that f is $\mathscr{M}$-measurable. □

Sometimes we wish to consider measurability on subsets of X.

Definition 29.7. If $(X, \mathscr{M})$ is a measurable space, f is a function on X, and $E \in \mathscr{M}$, we say that f is measurable on E if $f^{-1}(B) \cap E \in \mathscr{M}$ for all Borel sets B.

Definition 29.8. Let X be a set, $\{(Y_i, \mathscr{N}_i)\}_{i\in I}$ be a family of measurable spaces, and $f_i : X \to Y_i$. Define the σ-algebra *generated* by $\{f_i\}_{i\in I}$ as

$$\mathscr{M}\left(\{f_i^{-1}(E) \mid E_i \in \mathscr{N}_i,\, i \in I\}\right).$$

This is the smallest σ-algebra on X with respect to which the f_i's are all measurable.

In particular, if $X = \prod_{i\in I} Y_i$, we see that the product σ-algebra on X is the σ-algebra generated by the coordinate maps $\pi_i : X \to Y_i$.

Proposition 29.9. *Let $(X, \mathscr{M})$ and $(Y_i, \mathscr{N}_i)$ $(i \in I)$ be measurable spaces, $Y = \prod_{i\in I} Y_i$, $\mathscr{N} = \bigotimes_{i\in I} \mathscr{N}_i$, and $\pi_i : Y \to Y_i$ the coordinate maps. Then $f : X \to Y$ is $(\mathscr{M}, \mathscr{N})$-measurable iff $f_i = \pi_i \circ f$ is $(\mathscr{M}, \mathscr{N})$-measurable for all i.*

Proof.

$\boxed{\Longrightarrow}$ If f is measurable, so is each f_i since the composition of measurable maps is measurable.

$\boxed{\Longleftarrow}$ If each f_i is measurable, then for all $E_i \in \mathscr{N}_i$,

$$f_i^{-1}(E_i) = (\pi_i \circ f)^{-1}(E_i) = f^{-1}(\pi_i^{-1}(E_i)) \in \mathscr{M}.$$

By 29.3, f is measurable. □

Corollary 29.10. *A function $f : X \to \mathbb{C}$ is $\mathscr{M}$-measurable iff* Re f *and* Im f *are $\mathscr{M}$-measurable.*

Proof. This follows since $\mathscr{B}(\mathbb{C}) = \mathscr{B}(\mathbb{R}^2) = \mathscr{B}(\mathbb{R}) \otimes \mathscr{B}(\mathbb{R})$ by Proposition 1.5. □

We define Borel sets in the extended real number system $\overline{\mathbb{R}} = [-\infty, \infty]$ by

$$\mathscr{B}(\overline{\mathbb{R}}) := \{E \subset \overline{\mathbb{R}} \mid E \cap \mathbb{R} \in \mathscr{B}(\mathbb{R})\}.$$

We now establish that measurability is preserved under the familiar algebraic and limiting operations.

Proposition 29.11. *If $f, g : X \to \mathbb{C}$ are $\mathscr{M}$-measurable, then so are $f + g$ and fg.*

Proof. Define $F : X \to \mathbb{C} \times \mathbb{C}$ by $F(x) = (f(x), g(x))$. Since $\mathscr{B}(\mathbb{C} \times \mathbb{C}) = \mathscr{B}(\mathbb{C}) \otimes \mathscr{B}(\mathbb{C})$ by Proposition 1.5, F is $(\mathscr{M}, \mathscr{B}(\mathbb{C} \times \mathbb{C}))$-measurable by 2.4.

(i) Define $\phi : \mathbb{C} \times \mathbb{C} \to \mathbb{C}$ by $\phi(z, w) = z + w$. By 29.4, ϕ is $(\mathscr{B}(\mathbb{C} \times \mathbb{C}), \mathscr{B}(\mathbb{C}))$-measurable. Thus $f + g = \phi \circ F$ is $\mathscr{M}$-measurable.

(ii) Define $\psi\colon \mathbb{C}\times\mathbb{C}\to\mathbb{C}$ by $\psi(z,w)=zw$. By 29.4, ψ is $(\mathscr{B}(\mathbb{C}\times\mathbb{C}),\mathscr{B}(\mathbb{C}))$-measurable. Thus $fg=\psi\circ F$ is $\mathscr{M}$-measurable.

□

Suppose (f_n) is a sequence of extended-real functions on a set X. Define

$$\left(\sup_n f_n\right)(x) := \sup_n\left(f_n(x)\right),$$
$$\left(\limsup_{n\to\infty} f_n\right)(x) := \limsup_{n\to\infty}\left(f_n(x)\right).$$

If

$$f(x)=\lim_{n\to\infty} f_n(x) \quad (x\in X),$$

then we call f the *pointwise limit* of (f_n).

Proposition 29.12. *Let (f_n) be a sequence of $\overline{\mathbb{R}}$-valued measurable functions on $(X,\mathscr{M})$. Then*

$$\sup_n f_n,\quad \inf_n f_n,\quad \limsup_{n\to\infty} f_n,\quad \liminf_{n\to\infty} f_n$$

are all measurable. If $f(x)=\lim_{n\to\infty} f_n(x)$ exists for every $x\in X$, then f is measurable.

Proof. Let

$$g_1(x)=\sup_n f_n(x),\quad g_2(x)=\inf_n f_n(x),\quad g_3(x)=\limsup_{n\to\infty} f_n(x),\quad g_4(x)=\liminf_{n\to\infty} f_n(x).$$

We have

$$g_1^{-1}((a,\infty])=\bigcup_{n=1}^{\infty} f_n^{-1}((a,\infty]),\quad g_2^{-1}([-\infty,a))=\bigcup_{n=1}^{\infty} f_n^{-1}([-\infty,a)).$$

By Proposition 2.3, g_1 and g_2 are measurable.

More generally, if $h_k=\sup_{n>k} f_n$, then h_k is measurable for each k, so $g_3=\inf_k h_k$ is measurable, and likewise for g_4. Finally, if f exists then $f=g_3=g_4$, so f is measurable. □

Corollary 29.13. *If $f,g\colon X\to\overline{\mathbb{R}}$, then so are* $\max\{f,g\}$ *and* $\min\{f,g\}$.

Corollary 29.14. *If (f_n) is a sequence of complex-valued measurable functions and*

$$f(x)=\lim_{n\to\infty} f_n(x) \quad (x\in X)$$

then f is measurable.

Proof. Apply Corollary 2.5. □

Let $f\colon X \to \overline{\mathbb{R}}$. We define the *positive* and *negative parts* of f to be

$$f^+(x) := \max\{f(x), 0\}, \quad f^-(x) := \max\{-f(x), 0\}.$$

It is clear that

$$f = f^+ - f^- \quad \text{and} \quad |f| = f^+ + f^-$$

and it follows from these identities that

$$f^+ = \frac{1}{2}(|f| + f), \quad f^- = \frac{1}{2}(|f| - f).$$

Lemma. *If f is measurable, so are f^+ and f^-.*

Proof. By Corollary 2.8. □

Let $f\colon X \to \mathbb{C}$. We have its *polar decomposition*:

$$f = (\operatorname{sgn} f)|f|, \quad \text{where} \quad \operatorname{sgn} z = \begin{cases} z/|z| & (z \neq 0) \\ 0 & (z = 0) \end{cases}$$

Lemma. *If f is measurable, so are $|f|$ and $\operatorname{sgn} f$.*

Proof. $z \mapsto |z|$ is continuous on $\mathbb{C}$, and $z \mapsto \operatorname{sgn} z$ is continuous except at the origin. If $U \subset \mathbb{C}$ is open, $\operatorname{sgn}^{-1}(U)$ is either open or of the form $V \cup \{0\}$ where V is open, so sgn is Borel measurable. Therefore $|f| = |\cdot| \circ f$ and $\operatorname{sgn} f = \operatorname{sgn} \circ f$ are measurable. □

We now discuss the functions that are the building blocks for the theory of integration. Suppose $(X, \mathscr{M})$ is a measurable space. Let $E \subset X$. The characteristic function $\chi_E\colon X \to \mathbb{R}$ of E is defined by

$$\chi_E(x) := \begin{cases} 1 & (x \in E) \\ 0 & (x \notin E) \end{cases}$$

Lemma. *χ_E is measurable if and only if $E \in \mathscr{M}$.*

Proof.

$\boxed{\Longleftarrow}$ Suppose $E \in \mathscr{M}$. Let $B \subset \mathbb{R}$ be a Borel set. Then

$$\chi_E^{-1}(B) = \begin{cases} E & (0 \notin B, 1 \in B) \\ E^c & (0 \in B, 1 \notin B) \\ X & (0 \in B, 1 \in B) \\ \emptyset & (0 \notin B, 1 \notin B) \end{cases}$$

Since $E \in \mathscr{M}$, all the sets $E, E^c, X, \emptyset \in \mathscr{M}$. Hence χ_E is measurable.

$\boxed{\Longrightarrow}$ Suppose χ_E is measurable. Then for every Borel set $B \subset \mathbb{R}$, $\chi_E^{-1}(B) \in \mathscr{M}$. In particular, if $0 \notin B$ and $1 \in B$, then $\chi_E^{-1}(B) = E \in \mathscr{M}$. □

Definition 29.15 (Simple function)**.** We say f is a ***simple function*** on X if

$$f = \sum_{i=1}^{n} a_i \chi_{E_i}$$

for some $a_i \in \mathbb{C}$, $E_i \in \mathscr{M}$. We call this the *standard representation* of f.

It is clear that if f and g are simple functions, then so are $f+g$ and fg. We now show that arbitrary measurable functions can be approximated in a nice way by simple functions.

Theorem 29.16. *Let $(X, \mathscr{M})$ be a measurable space.*

(i)

(ii)

If μ is a measure on $(X, \mathscr{M})$, we may wish to except μ-null sets from consideration in studying measurable functions. In this respect, life is a bit simpler if μ is complete.

Proposition 29.17. *The following implications are valid iff the measure μ is complete:*

(i) If f is measurable and $f = g$ μ-a.e., then g is measurable.

(ii) If f_n is measurable for $n \in \mathbb{N}$ and $f_n \to f$ μ-a.e., then f is measurable.

Proposition 29.18. *Let $(X, \mathscr{M}, \mu)$ be a measure space, and let $(X, \overline{\mathscr{M}}, \overline{\mu})$ be its completion. If f is an $\overline{\mathscr{M}}$-measurable function on X, then there exists an $\overline{\mathscr{M}}$-measurable function g such that $f = g$ $\overline{\mu}$-almost everywhere.*

29.2 Integration of Non-negative Functions

Fix a measure space $(X, \mathscr{M}, \mu)$. Define

$$L^+ = \text{the space of all measurable functions from } X \text{ to } [0,\infty].$$

Definition 29.19. Let $\phi \in L^+$ be a simple function with standard representation $\phi = \sum_{i=1}^n a_i \chi_{E_i}$. Define the ***integral*** of ϕ with respect to μ by

$$\int \phi \, \mathrm{d}\mu := \sum_{i=1}^n a_i \mu(E_i).$$

Notation. When there is no danger of confusion, we shall also write

$$\int \phi = \int \phi \, \mathrm{d}\mu = \int_X \phi \, \mathrm{d}\mu = \int \phi(x) \, \mathrm{d}\mu(x).$$

Instead of integrating over the whole space X, we can integrate over a subset. If $A \in \mathscr{M}$, then $\phi \chi_A$ is also simple (let $\phi = \sum_{i=1}^n a_i \phi_{A_i}$, then $\phi \chi_A = \sum_{i=1}^n a_i \chi_{A_i \cap A}$). Thus we define

$$\int_A \phi \, \mathrm{d}\mu := \int \phi \chi_A \, \mathrm{d}\mu.$$

The next result summarises basic properties of the integrals of simple functions.

Proposition 29.20. *Let ϕ and ψ be simple functions in L^+.*

(i) If $c \geq 0$, $\int c\phi = c \int \phi$. *(scalar multiplication)*

(ii) $\int (\phi + \psi) = \int \phi + \int \psi$. *(addition)*

(iii) If $\phi \leq \psi$, then $\int \phi \leq \int \psi$. *(monotonicity)*

(iv) The map $A \mapsto \int_A \mathrm{d}\mu$ is a measure on $\mathscr{M}$.

Proof.

(i) Let $\phi = \sum_{i=1}^n a_i \chi_{E_i}$ be the standard representation. Then

$$\int c\phi = \sum_{i=1}^n c a_i \mu(E_i) = c \sum_{i=1}^n a_i \mu(E_i) = c \int \phi.$$

(ii) Let $\phi = \sum_{i=1}^n a_i \chi_{E_i}$ and $\psi = \sum_{j=1}^m b_j \chi_{F_j}$ be the standard representations. Then $E_i = \bigcup_{j=1}^m (E_i \cap F_j)$ and $F_j = \bigcup_{i=1}^n (E_i \cap F_j)$ since $\bigcup_{i=1}^n E_i = \bigcup_{j=1}^m F_j = X$, and these unions are

disjoint. Hence the finite additivity of μ implies that

$$\int \phi + \int \psi = \sum_{i,j} (a_i + b_j)\mu(E_i \cap F_j),$$

and the same reasoning shows that the sum on the right equals $\int (\phi + \psi)$.

(iii) If $\phi \leq \psi$, then $a_i \leq b_j$ whenever $E_i \cap F_j \neq \emptyset$, so

$$\int \phi = \sum_{i,j} a_i \mu(E_i \cap F_j) \leq \sum_{i,j} b_j \mu(E_i \cap F_j) = \int \psi.$$

(iv)

□

Previously we defined the integral for simple functions in L^+. We now extend the integral to *all* functions in L^+.

Definition 29.21 (Integral). Let $f \in L^+$. Define the ***integral*** of f with respect to μ by

$$\int f \, d\mu := \sup \int \phi \, d\mu$$

where the supremum is taken over all simple functions ϕ, with $0 \leq \phi \leq f$.

Remark. Since $\phi = 0$ is simple, the above definition makes sense (we are not taking the supremum over an empty set).

If f is simple, the two definitions of $\int f$ agree, because $\phi \leq f \implies \int \phi \leq \int f$, and the family of simple functions over which the supremum is taken includes f itself.

From the definition, we can immediately deduce the following properties:

(i) If $f \leq g$, then $\int f \leq \int g$. (monotonicity)

(ii) If $c \geq 0$, $\int cf = c \int f$. (scalar multiplication)

We now establish one of the fundamental convergence theorems.

Theorem 29.22 (Monotone convergence theorem). *Let (f_n) be a sequence in L^+. Suppose $f_n \to f$ pointwise, and $f_n \leq f_{n+1}$ for all n. Then $f \in L^+$, and*

$$\int f = \lim_{n\to\infty} \int f_n. \tag{29.1}$$

Proof. We will prove (i) $\lim_{n\to\infty} \int f_n \leq \int f$, and (ii) $\lim_{n\to\infty} \int f_n \geq \int f$.

(i) By monotonicity, $f_n \le f_{n+1} \implies \int f_n \le \int f_{n+1}$ for all n, so $(\int f_n)$ is an increasing sequence of numbers; thus its limit exists (possibly equal to ∞).

Moreover, $f_n \le f \implies \int f_n \le \int f$ for all n. Hence $\lim_{n\to\infty} \int f_n \le \int f$.

(ii) Take any simple function $\phi : X \to [0,\infty]$ with $0 \le \phi \le f$. Fix $\alpha \in (0,1)$. Define

$$E_n := \{x \in X \mid f_n(x) \ge \alpha\phi(x)\}.$$

Since $f_n \le f_{n+1}$ for all n, we have $E_n \subset E_{n+1}$ for all n; thus (E_n) is an increasing sequence of measurable sets: $E_1 \subset E_2 \subset \cdots$. Since $f_n \to f$, we have $\bigcup_{n=1}^{\infty} E_n = X$. Since $A \mapsto \int_A \phi \, d\mu$ is a measure by 29.20, we have

$$\int \phi \, d\mu = \lim_{n\to\infty} \int_{E_n} \phi \, d\mu$$

by continuity from below (see 28.16). However

$$\alpha \int_{E_n} \phi \, d\mu = \int_{E_n} \alpha\phi \, d\mu \le \int_{E_n} f_n \, d\mu \le \int f_n \, d\mu \,.$$

Hence

$$\lim_{n\to\infty} \int f_n \ge \alpha \lim_{n\to\infty} \int_{E_n} \phi = \alpha \int \phi.$$

Since this is true for all $\alpha < 1$, it remains true for $\alpha = 1$; thus

$$\lim_{n\to\infty} \int f_n \ge \int \phi.$$

Taking the supremum of the RHS over all simple $\phi \le f$, we obtain $\lim_{n\to\infty} f_n \ge \int f$.

□

The condition that (f_n) is increasing is required. **Counterexample**: define f_n such that it traces out a triangle with base $[0, 2/n]$ and height n. Then $\int f_n = 1$, but $f_n \to 0$ implies $\int 0 = 0$.

The definition of $\int f$ involves the supremum over a huge (usually uncountable) family of simple functions, so it may be difficult to evaluate $\int f$ directly from the definition. The monotone convergence theorem, however, assures us that to compute $\int f$ it is enough to compute $\lim_{n\to\infty} \int \phi_n$ where (ϕ_n) is any sequence of simple functions that increase to f, and Theorem 2.10 guarantees that such sequences exist.

As a first application, we establish the additivity of the integral:

Theorem 29.23. *If (f_n) is a finite or infinite sequence in L^+, and $f = \sum_n f_n$, then*

$$\int f = \sum_n \int f_n.$$

Proposition 29.24. *If $f \in L^+$, then $\int f = 0 \iff f = 0$ a.e.*

Corollary 29.25. *If (f_n) is a sequence in L^+, $f \in L^+$, and $f_n(x)$ increases to $f(x)$ for a.e. x, then*

$$\int f = \lim_{n\to\infty} \int f_n.$$

Proof. Suppose $f_n(x)$ increases to $f(x)$ for $x \in E$, where $\mu(E^c) = 0$. Then $f - f\chi_E = 0$ a.e. and $f_n - f_n\chi_E = 0$ a.e.. Hence by the monotone convergence theorem,

$$\int f = \int f\chi_E = \lim_{n\to\infty} \int f_n\chi_E = \lim_{n\to\infty} \int f_n.$$

□

Theorem 29.26 (Fatou's lemma)**.** *If (f_n) is a sequence in L^+, then*

$$\int \liminf_{n\to\infty} f_n \le \liminf_{n\to\infty} \int f_n.$$

Proof. For each $k \ge 1$ we have $\inf_{n\ge k} f_n \le f_j$ for $j \ge k$, hence $\int \inf_{n\ge k} f_n \le \int f_j$ for $j \ge k$, hence $\int \inf_{n\ge k} f_n \le \inf_{j\ge k} \int f_j$. Now let $k \to \infty$ and apply the monotone convergence theorem:

$$\int \left(\liminf_{n\to\infty} f_n\right) = \int \lim_{k\to\infty} \left(\inf_{n\ge k} f_n\right) = \lim_{k\to\infty} \int \left(\inf_{n\ge k} f_n\right) \le \lim_{n\to\infty} \int f_n.$$

□

Corollary 29.27. *If (f_n) is a sequence in L^+, $f \in L^+$, and $f_n \to f$ a.e., then*

$$\int f \le \liminf_{n\to\infty} \int f_n.$$

Proposition 29.28. *If $f \in L^+$ and $\int f < \infty$, then*

(i) $\{x \in X \mid f(x) = \infty\}$ *is a null set;*

(ii) $\{x \in X \mid f(x) > 0\}$ *is σ-finite.*

29.3 Integration of Complex Functions

Fix a measure space $(X, \mathscr{M}, \mu)$. The integral defined in the previous section can be extended to real-valued measurable functions f in an obvious way.

Let f^+ and f^- denote the positive and negative parts of f respectively.

Definition 29.29. Suppose at least one of $\int f^+$ and $\int f^-$ are finite. We define

$$\int f := \int f^+ - \int f^-.$$

We shall be mainly concerned with the case where $\int f^+$ and $\int f^-$ are both finite; we then say that f is ***integrable***.

Since $|f| = f^+ + f^-$, it is clear that f is integrable iff $\int |f| < \infty$.

Proposition 29.30. *The set of integrable real-valued functions on X is a real vector space, and the integral is a linear functional on it.*

Definition 29.31. If f is a complex-valued measurable function, we say f is ***integrable*** if

$$\int |f| < \infty.$$

More generally, if $E \in \mathscr{M}$, f is ***integrable on*** E if

$$\int_E |f| < \infty.$$

Since $|f| \leq |\operatorname{Re} f| + |\operatorname{Im} f| \leq 2|f|$, f is integrable iff $\operatorname{Re} f$ and $\operatorname{Im} f$ are both integrable, and in this case we define

$$\int f = \int \operatorname{Re} f + i \int \operatorname{Im} f.$$

It follows easily that the space of complex-valued integrable functions is a complex vector space and that the integral is a complex-linear functional on it. We denote this space by $L^1(\mu)$, or simply L^1.

Proposition 29.32 (Triangle inequality). *If $f \in L^1$, then*

$$\left| \int f \right| \leq \int |f|.$$

We now present the last of the three basic convergence theorems (the other two being the monotone convergence theorem and Fatou's lemma) and derive some useful consequences from

it.

Theorem 29.33 (Dominated convergence theorem). *Suppose (f_n) is a sequence in L^1 such that*

(i) $f_n \to f$ a.e.

(ii) there exists a non-negative $g \in L^1$ such that $|f_n| \le g$ a.e. for all n.

Then $f \in L^1$, and

$$\int f = \lim_{n\to\infty} \int f_n. \tag{29.2}$$

Theorem 29.34. *Suppose (f_n) is a sequence in L^1 such that $\sum_{n=1}^{\infty} \int |f_n| < \infty$. Then $\sum_{n=1}^{\infty} f_n$ converges a.e. to a function in L^1, and*

$$\int \sum_{n=1}^{\infty} f_n = \sum_{n=1}^{\infty} \int f_n.$$

If we let μ be the counting measure on a countable set, Theorem 1.27 is a statement about double series of nonnegative real numbers (which can of course be proved by more elementary means):

Corollary 29.35. *If $a_{ij} \ge 0$ for $i, j = 1, 2, \ldots$, then*

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} a_{ij}.$$

Theorem 29.36 (Fatau's lemma). *Let $f_n \colon X \to [0,\infty]$ be a sequence of measurable functions. Then*

$$\int_X \left(\liminf_{n\to\infty} f_n\right) \mathrm{d}\mu \le \liminf_{n\to\infty} \int_X f_n \,\mathrm{d}\mu\,. \tag{29.3}$$

Proposition 29.37 (Change of variables). *Suppose $f \colon X \to [0,\infty]$ is measurable, and*

$$\phi(E) = \int_E f \,\mathrm{d}\mu \quad (E \in \mathscr{M}.)$$

Then ϕ is a measure on $\mathscr{M}$, and

$$\int_X g \,\mathrm{d}\phi = \int_X gf \,\mathrm{d}\mu \tag{29.4}$$

for every measurable $g \colon X \to [0,\infty]$.

Lebesgue Integral

29.4 Modes of Convergence

Another mode of convergence that is frequently useful is convergence in measure.

Definition 29.38. We say a sequence (f_n) of measurable complex-valued functions on $(X, \mathscr{M}, \mu)$ is ***Cauchy in measure*** if

$$\forall \varepsilon > 0, \quad \exists N \in \mathbb{N}, \quad \forall m, n \geq N, \quad \mu(\{x : |f_n(x) - f_m(x)| \geq \varepsilon\}) < \varepsilon.$$

Definition 29.39. We say a sequence (f_n) of measurable complex-valued functions on $(X, \mathscr{M}, \mu)$ ***converges in measure*** to f if

$$\forall \varepsilon > 0, \quad \exists N \in \mathbb{N}, \quad \forall n \geq N, \quad \mu(\{x : |f_n(x) - f(x)| \geq \varepsilon\}) < \varepsilon.$$

Proposition 29.40. *If $f_n \to f$ in L^1, then $f_n \to f$ in measure.*

The converse is false.

Theorem 29.41. *Suppose (f_n) is Cauchy in measure. Then there exists a measurable function f such that $f_n \to f$ in measure, and there exists a subsequence (f_{n_k}) that converges to f a.e. Moreover, if also $f_n \to g$ in measure, then $g = f$ a.e.*

Corollary 29.42. *If $f_n \to f$ in L^1, there exists a subsequence (f_{n_k}) such that $f_{n_k} \to f$ a.e.*

Theorem 29.43 (Egorov's theorem)**.** *Suppose $(X, \mathscr{M}, \mu)$ is a measure space with $\mu(X) < \infty$. Suppose (f_n) is a sequence of measurable functions from X to $\mathbb{R}$, and $f_n \to f$ on X. Then*

$$\forall \varepsilon > 0, \quad \exists E \in \mathscr{M}, \quad \mu(E) < \varepsilon \quad \text{and} \quad f_n \rightrightarrows f \text{ on } E^c.$$

The type of convergence involved in the conclusion of Egorov's theorem is sometimes called *almost uniform convergence*. It is not hard to see that almost uniform convergence implies a.e. convergence and convergence in measure.

29.5 Product Measures

29.6 The n-dimensional Lebesgue Integral

29.7 Integration in Polar Coordinates

IX
Graph Theory

A very early theorem of *graph theory*, perhaps even the first, was proved in 1766 by Euler, concerning a popular problem of the time called "the bridges of Königsberg". Königsberg is divided into 4 districts by the river Pregel and has 7 bridges. The problem was to decide whether it is possible to take a walk that crosses every bridge exactly once. To formulate this problem mathematically, we construct a graph in which there is a vertex for each district and an edge representing each bridge.

30 The Basics

30.1 Graphs

By $[A]^k$ we denote the set of all k-element subsets of A. Sets with k elements will be called *k-sets*; subsets with k elements are *k-subsets*.

Definition 30.1 (Graph)**.** A ***graph*** is a pair $G = (V, E)$ of sets such that $E \subset [V]^2$. The elements of V are the *vertices*; the elements of E are the *edges*.

A graph can be represented visually by drawing a point for each vertex, and a line between any pair of points that form an edge.

A graph with vertex set V is said to be a graph *on* V. The vertex set of a graph G is referred to as $V(G)$, its edge set as $E(G)$.

Notation. We shall not always distinguish strictly between a graph and its vertex or edge set. For example, we may speak of a vertex $v \in G$ (rather than $v \in V(G)$), an edge $e \in G$, and so on.

Definition 30.2 (Order)**.** The number of vertices of a graph G is its ***order***, written as $|G|$; its number of edges is denoted by $\|G\|$.

Graphs are *finite*, *infinite*, *countable* and so on according to their order.

Unless stated otherwise, our graphs will be finite.

The *empty graph* $(\emptyset, \emptyset)$ is denoted $\emptyset$; a *trivial graph* is a graph of order 0 or 1.

Remark. Sometimes, e.g., to start an induction, trivial graphs can be useful; at other times they form silly counterexamples and become a nuisance. To avoid cluttering the text with non-triviality conditions, we shall mostly treat the trivial graphs, and particularly the empty graph $\emptyset$, with generous disregard.

Definition 30.3. A vertex v is ***incident*** with an edge e if $v \in e$; then e is an edge *at* v. The two vertices incident with an edge are its *endvertices* or *ends*, and an edge *joins* its ends.

An edge $\{x,y\}$ is usually written as xy (or yx). If $x \in X$ and $y \in Y$, then xy is an $X-Y$ edge. The set of all $X-Y$ edges in a set E is denoted by $E(X,Y)$; instead of $E(\{x\},Y)$ and $E(X,\{y\})$ we simply write $E(x,Y)$ and $E(X,y)$. The set of all the edges in E at a vertex v is denoted by $E(v)$.

Definition 30.4. Two vertices x,y of G are ***adjacent***, or *neighbours*, if xy is an edge of G.
Two edges $e \neq f$ are ***adjacent*** if they have an end in common.

If all neighbour the vertices of G are pairwise adjacent, then G is ***complete***. A complete graph on n vertices is denoted K_n.

Pairwise non-adjacent vertices or edges are called *independent*. More formally, a set of vertices or of edges is *independent* if no two of its elements are adjacent. Independent sets of vertices are also called *stable*.

Definition 30.5 (Homomorphism)**.** Let $G = (V,E)$ and $G' = (V',E')$ be two graphs. A map $\phi : V \to V'$ is a ***homomorphism*** from G to G' if it preserves the adjacency of vertices:

$$xy \in E \implies \phi(x)\phi(y) \in E'.$$

Then, by considering the contrapositive of the above, for every vertex x' in the image of ϕ, its inverse image $\phi^{-1}(x')$ is an independent set of vertices in G.

Definition 30.6 (Isomorphism)**.** If ϕ is bijective and its inverse ϕ^{-1} is also a homomorphism:

$$\forall x,y \in V, \quad xy \in E \iff \phi(x)\phi(y) \in E'$$

we call ϕ an ***isomorphism***, say that G and G' are ***isomorphic***, and write $G \cong G'$.

An isomorphism from G to itself is an *automorphism* of G.

Remark. We do not normally distinguish between isomorphic graphs. Thus, we usually write $G = G'$ rather than $G \cong G'$, speak of the complete graph on 17 vertices, and so on.

A class of graphs that is closed under isomorphism is called a *graph property*. For example, "containing a triangle" is a graph property: if G contains three pairwise adjacent vertices then so does every graph isomorphic to G.

A map taking graphs as arguments is called a *graph invariant* if it assigns equal values to isomorphic graphs. The number of vertices and the number of edges of a graph are two simple graph invariants; the greatest number of pairwise adjacent vertices is another.

Define the *union* and *intersection* of two graphs $G = (V,E)$ and $G' = (V',E')$ as

$$G \cup G' := (V \cup V', E \cup E'),$$
$$G \cap G' := (V \cap V', E \cap E').$$

If $G \cap G' = \emptyset$, then G and G' are *disjoint*.

Definition 30.7 (Subgraph)**.** We say G' is a ***subgraph*** of G, denoted by $G' \subset G$, if $V' \subset V$ and $E' \subset E$.

Less formally, we say that G *contains* G'. If $G' \subset G$ and $G' \neq G$, then G' is a *proper subgraph* of G.

Furthermore, if $G' \subset G$, we say

- G' is an ***induced subgraph*** if G' contains all the edges $xy \in E$ with $x, y \in V'$; we say that V' *induces* or *spans* G' in G, and write $G' =: G[V']$.

  Thus if $U \subset V$ is any set of vertices, then $G[U]$ denotes the graph on U whose edges are precisely the edges of G with both ends in U. If H is a subgraph of G, not necessarily induced, we abbreviate $G[V(H)]$ to $G[H]$.

- G' is a ***spanning subgraph*** of G if V' spans all of G, i.e., $V' = V$.

If U is any set of vertices (usually of G), we write $G - U$ for $G[V \setminus U]$. In other words, $G - U$ is obtained from G by deleting all the vertices in $U \cap V$ and their incident edges. If $U = \{v\}$ is a singleton, we write $G - v$ rather than $G - \{v\}$. Instead of $G - V(G')$ we simply write $G - G'$. For a subset F of $[V]^2$ we write

$$G - F := (V, E \setminus F)$$
$$G + F := (V, E \cup F);$$

as above, $G - \{e\}$ and $G + \{e\}$ are abbreviated to $G - e$ and $G + e$.

We call G ***edge-maximal*** with a given graph property if G itself has the property but no graph (V, F) with $F \supsetneq E$ does.

Remark. More generally, when we call a graph *minimal* or *maximal* with some property but have not specified any particular ordering, we are referring to the subgraph relation. When we speak of minimal or maximal sets of vertices or edges, the reference is simply to set inclusion.

If G and G' are disjoint, we denote by $G * G'$ the graph obtained from $G \cup G'$ by joining all the vertices of G to all the vertices of G'.

Example. $K_2 * K_3 = K_5$.

Definition 30.8 (Complement)**.** The ***complement*** $\overline{G}$ of G is the graph on V with edge set $[V]^2 \setminus E$:

$$xy \in E(\overline{G}) \iff xy \notin E(G).$$

That is, $\overline{G}$ has edges exactly where there are no edges in G.

Definition 30.9 (Line graph)**.** The ***line graph*** $L(G)$ of G is the graph on E in which $x, y \in E$ are adjacent as vertices if and only if they are adjacent as edges in G.

30.2 The Degree of a Vertex

Let $G = (V, E)$ be a (non-empty) graph.

Let the set of neighbours of a vertex $v \in G$ be denoted by $N_G(v)$, or simply $N(v)$. More generally, for $U \subset V$, the neighbours in $V \setminus U$ of vertices in U are called *neighbours* of U; their set is denoted by $N(U)$.

Definition 30.10 (Degree). The ***degree*** $d(v)$ of a vertex v is

$$d(v) := |E(v)| = \text{number of neighbours of } v.$$

A vertex of degree 0 is said to be *isolated*; a vertex of degree 1 is called a *leaf*.

The *minimum degree* and *maximum degree* of G are defined respectively as

$$\delta(G) := \min_{v \in V} d(v),$$
$$\Delta(G) := \max_{v \in V} d(v).$$

If all the vertices of G have the same degree k, then G is *k-regular*, or simply *regular*. In particular, a 3-regular graph is called *cubic*.

The *average degree* of G is

$$d(G) := \frac{1}{|V|} \sum_{v \in V} d(v).$$

Clearly,

$$\delta(G) \leq d(G) \leq \Delta(G).$$

The average degree quantifies globally what is measured locally by the vertex degrees: the number of edges of G per vertex. Sometimes it will be convenient to express this ratio directly, as

$$\varepsilon(G) := \frac{|E|}{|V|}.$$

The quantities d and ε are, of course, intimately related. Indeed, if we sum up all the vertex degrees in G, we count every edge exactly twice: once from each of its ends. Thus

$$|E| = \frac{1}{2} \sum_{v \in V} d(v) = \frac{1}{2} d(G) \cdot |V|.$$

(The first equality is called the *degree sum formula*). Therefore

$$\varepsilon(G) = \frac{1}{2} d(G).$$

Proposition 30.11 (Handshaking lemma)**.** *The number of vertices of odd degree in a graph is always even.*

Proof. The degree sum formula states that

$$\sum_{v \in V} d(v) = 2|E|.$$

The sum of vertices can be written as

$$\sum_{v \in V} d(v) = \sum_{\substack{v \in V \\ d(v) \text{ even}}} d(v) + \sum_{\substack{v \in V \\ d(v) \text{ odd}}} d(v).$$

Now $\sum_{v \in V} d(v)$ and $\sum_{\substack{v \in V \\ d(v) \text{ even}}} d(v)$ are even, so $\sum_{\substack{v \in V \\ d(v) \text{ odd}}} d(v)$ is even. Hence the number of odd-degree vertices must be even. □

If a graph has large minimum degree, i.e., everywhere, locally, many edges per vertex, it also has many edges per vertex globally: $\varepsilon(G) = \frac{1}{2}d(G) \geq \frac{1}{2}\delta(G)$. Conversely, of course, its average degree may be large even when its minimum degree is small. However, the vertices of large degree cannot be scattered completely among vertices of small degree: as the next proposition shows, every graph G has a subgraph whose average degree is no less than the average degree of G, and whose minimum degree is more than half its average degree:

Proposition 30.12. *Every graph G with at least one edge has a subgraph H with $\delta(H) > \varepsilon(H) \geq \varepsilon(G)$.*

Proof. To construct H from G, let us try to delete vertices of small degree one by one, until only vertices of large degree remain. Up to which degree $d(v)$ can we afford to delete a vertex v, without lowering ε? Clearly, up to $d(v) = \varepsilon$: then the number of vertices decreases by 1 and the number of edges by at most ε, so the overall ratio ε of edges to vertices will not decrease.

Formally, we construct a sequence $G \supset G_0 \supset G_1 \supset \cdots$ of induced subgraphs of G as follows. If G_i has a vertex v_i of degree $d(v_i) \leq \varepsilon(G_i)$, we let $G_{i+1} := G_i - v_i$; if not, we terminate our sequence and set $H := G_i$. By the choices of v_i, we have $\varepsilon(G_{i+1}) \geq \varepsilon(G_i)$ for all i, and hence $\varepsilon(H) \geq \varepsilon(G)$.

What else can we say about the graph H? Since $\varepsilon(K_1) = 0 < \varepsilon(G)$, none of the graphs in our sequence is trivial, so in particular $H \neq \emptyset$. The fact that H has no vertex suitable for deletion thus implies $\delta(H) > \varepsilon(H)$, as claimed. □

30.3 Paths and Cycles

Definition 30.13 (Walk). A $u-v$ ***walk*** in G, denoted by W, is a finite sequence of vertices

$$u = v_0, v_1, \ldots, v_n = v$$

such that $v_i v_{i+1} \in E$ for all $0 \leq i \leq n-1$.

- If $u = v$, we call W a *closed walk*; otherwise, it is an *open walk*.
- If the vertices in W are distinct, we call it a ***path***.
- A ***circuit*** is a closed walk without repeating edges, i.e. $u = v$, it begins and ends with the same vertex.
- A ***cycle*** is a closed walk without repeating vertices, other than the initial and terminal vertices, i.e. $u = v$ but the vertices are otherwise distinct and W has at least 3 vertices. If a graph G has no cycle we call it ***acyclic***.
- A ***trail*** is a walk in which no two vertices appear consecutively (in either order) more than once; that is, no edge is used more than once. A ***tour*** is a closed trail.

Proposition 30.14. *Every graph G contains a path of length $\delta(G)$ and a cycle of length at least $\delta(G)+1$ (provided that $\delta(G) \geq 2$).*

The ***distance*** $d(x,y)$ in G of two vertices x, y is the length of a shortest $x-y$ path in G; if no such path exists, we set $d(x,y) := \infty$.

Definition 30.15 (Diameter). The ***diameter*** of a graph G is the greatest distance between any two vertices in G:

$$\operatorname{diam} G := \max_{x \neq y} d(x,y).$$

By construction, $d(x,y) \leq \operatorname{diam} G$ for all $x, y \in V$.

Diameter and girth are, of course, related:

Proposition 30.16. *Every graph G containing a cycle satisfies $g(G) \leq 2\operatorname{diam}(G)+1$.*

Proposition 30.17. *A graph G of radius at most k and maximum degree at most $d \geq 3$ has fewer than $\frac{d}{d-2}(d-1)^k$ vertices.*

Theorem 30.18. *Let G be a graph. If $d(G) \geq d \geq 2$ and $g(G) \geq g \in \mathbb{N}$, then $|G| \geq n_0(d,g)$.*

Corollary 30.19. *If $\delta(G) \geq 3$ then $g(G) < 2\log|G|$.*

30.4 Connectivity

Definition 30.20. A graph G is ***connected*** if for all $x, y \in V$, there exists a $x - y$ path. If $U \subset V(G)$ and $G[U]$ is connected, we also call U itself connected (in G).

Proposition 30.21. *The vertices of a connected graph G can always be enumerated, say as $v_1, \ldots, v_n$, so that $G_i := G[v_1, \ldots, v_i]$ is connected for every i.*

30.5 Trees and Forests

Definition 30.22 (Tree)**.** A ***tree*** is a connected graph that does not contain any cycles; that is, it is a minimally connected graph.

Proposition 30.23. *Any tree is acyclic.*

Proof. Let G be a tree, i.e. G is minimally connected.

Suppose, for a contradiction, that G contains a cycle C. Let $e \in E(C)$. We will obtain our contradiction by showing that $G - e := (V(G), E(G) \setminus \{e\})$ is connected.

Let P be the path obtained by deleting e from C. Consider any u, v in $V(G)$. As G is connected, there is an $u - v$ walk W in G. Replacing any use of e in W by P gives an $u - v$ walk in $G - e$. Thus $G - e$ is connected, a contradiction. □

The following are equivalent characterisations of trees.

Lemma 30.24 (Characterisation of trees)**.**

(i) G is a tree if and only if G is connected and acyclic.

(ii) Any two vertices in a tree are joined by a unique path.

Proof.

(i) If G is a tree then G is connected and acyclic.

Conversely, let G be connected and acyclic. Suppose for a contradiction that $G - e$ is connected for some $e = (u, v) \in E(G)$.

Let W be a shortest $u - v$ walk in $G - e$. Then W must be a path, i.e. have no repeated vertices, otherwise we would find a shorter walk by deleting a segment of W between two visits to the same vertex. Combining W with (u, v) gives a cycle, which is a contradiction.

(ii) Suppose for a contradiction that this fails for some tree G.

Choose u, v in $V(G)$ so that there are distinct $u - v$ paths P1, P2, and P1 is as short as possible over all such choices of u and v.

Then P1 and P2 only intersect in u and v, so their union is a cycle, contradicting Proposition 30.23.

□

Remark. The fact that a shortest walk between two points is a path is often useful. More generally, considering an extremal (shortest, longest, minimal, maximal, ...) object is often a useful proof technique.

Lemma 30.25. *Any tree with at least two vertices has at least two leaves.*

Proof. Consider any tree G. Let P be a longest path in G. The two ends of P must be leaves. Indeed, an end cannot have a neighbour in $V(G) \setminus V(P)$, or we could make P longer, and cannot have any neighbour in $V(P)$ other than the next in the sequence of P, or we would have a cycle.

The existence of leaves in trees is useful for inductive arguments, via the following lemma. Given $v \in V(G)$, let $G-v$ be the graph with $V(G-v) = V(G) \setminus \{v\}$ and $E(G-v) = \{(u,v) \in E(G) \mid v \notin \{u,v\}\}$. □

Lemma 30.26. *If G is a tree and v is a leaf of G then $G-v$ is a tree.*

Proof. By Lemma 30.24 (i), it suffices to show that $G-v$ is connected and acyclic. Acyclicity is immediate from Proposition 30.23. Connectedness follows by noting for any $u,v \in V(G) \setminus \{v\}$ that the unique $u-v$ path in G is contained in $G-v$. □

Lemma 30.27. *Any tree on n vertices has $n-1$ edges.*

Proof. By induction. A tree with 1 vertex has 0 edges. Let G be a tree on $n > 1$ vertices. By Lemma 30.25, G has a leaf v. By Lemma 30.26, $G-v$ is a tree. By induction hypothesis, $G-v$ has $n-2$ edges. Replacing v gives $n-1$ edges in G. □

We conclude this section with another characterisation of trees. First we note that any connected graph G contains a minimally connected subgraph (i.e. a tree) with the same vertex set, which we call a ***spanning tree*** of G.

Lemma 30.28. *A graph G is a tree on n vertices if and only if G is connected and has $n-1$ edges.*

Proof. If G is a tree then G is connected by definition and has $n-1$ edges by Lemma 30.27.

Conversely, suppose that G is connected and has $n-1$ edges. Let H be a spanning tree of G. Then H has $n-1$ edges by Lemma 30.27, so $H = G$, so G is a tree. □

30.6 Bipartite Graphs

Definition 30.29. Let $r \geq 2$ be an integer. A graph $G = (V,E)$ is called ***r-partite*** if V admits a partition into r classes such that each edge has its ends in different classes: vertices in the same partition class must not be adjacent.

Instead of "2-partite", one usually says ***bipartite***.

An r-partite graph in which every two vertices from different partition classes are adjacent is called ***complete***; the complete r-partite graphs for all r together are the *complete multipartite graphs*. The complete r-partite graph $\overline{K_{n_1}} * \cdots * \overline{K_{n_r}}$ is denoted by $K_{n_1,\ldots,n_r}$; if $n_1 = \cdots = n_r =: s$, we abbreviate this to K_r^s. Thus K_r^s is the complete r-partite graph in which every partition class contains exactly s vertices.

Clearly, a bipartite graph cannot contain an odd cycle, a cycle of odd length. In fact, the bipartite graphs are characterised by this property:

Proposition 30.30. *A graph is bipartite if and only if it contains no odd cycle.*

30.7 Contraction and Minors

Proposition 30.31. *The minor relation $\preceq$ and the topological-minor relation are partial orderings on the class of finite graphs.*

Corollary 30.32.

Proposition 30.33.

30.8 Euler Tours

Definition 30.34 (Euler tour)**.** An ***Euler tour*** is a closed walk that traverses every edge of the graph exactly once.
A graph is ***Eulerian*** if it admits an Euler tour.

Theorem 30.35 (Euler)**.** *A connected graph is Eulerian if and only if every vertex has even degree.*

Proof.

$\boxed{\Longrightarrow}$ A vertex appearing k times in an Euler tour (or $k+1$ times, if it is the starting and finishing vertex and as such counted twice) must have degree $2k$.

Hence the degree of every vertex is even.

$\boxed{\Longleftarrow}$ We want to show that every connected graph G with all degrees even has an Euler tour.

Induct on the number of edges $\|G\|$. For the base case $\|G\| = 0$, there are no edges and one vertex, so G is trivially Eulerian.

Now suppose $\|G\| \geq 1$, and the desired result holds for all connected graphs with fewer than $\|G\|$ edges. Let G be a connected graph where every vertex has even degree.

Since all degrees are even, we can find a non-trivial closed walk that contains no edge more than once:

- Start walking from any vertex.
- At each step, there is an unused edge to continue because degree is even.
- Eventually you must return to a previously visited vertex. This gives a closed walk.

Let W be such a closed walk of *maximal* length; let F denote the set of its edges. If $F = E(G)$, then W is an Euler tour, so we are done. Otherwise, there are edges left over. Let $G' := G - F$, the subgraph with the remaining edges.

Every vertex in G has even degree. Since each edge in F connects two vertices, removing those edges keeps the degree even at every vertex in G'. Thus all degrees in G' are even.

Since G is connected, G' has an edge e incident with a vertex on W. Let C be the connected component of G' containing e. By induction hypothesis, C has an Euler tour, say W'.

Since W' shares at least one vertex with W, we can *insert* W' into W at the shared vertex (suitable re-indexed). This yields a longer closed walk in G, contradicting the maximal length of W. □

An Euler tour can be found efficiently using *Fleury's Algorithm*:

Start at any vertex. We will follow a walk, erasing each edge after it is used (erased edges cannot be used again). At each stage, ensure that the following holds:

(i) when the edge is removed, the resulting graph is connected once isolated vertices are removed, and

(ii) we do not run along an edge to a leaf, unless this is the only edge of the graph.

30.9 Some Linear Algebra

How do we specify a graph? We can list the vertices and edges, but there are other useful representations.

Definition 30.36. The ***adjacency matrix*** $A = (a_{ij})_{n\times n}$ of G is defined by

$$a_{ij} := \begin{cases} 1 & (v_i v_j \in E) \\ 0 & (\text{otherwise}) \end{cases}$$

30.10 Other Notions of Graphs

Definition 30.37. A ***loop*** is an edge vv for some $v \in V(G)$; that is, an edge whose endpoints are equal. An edge is a ***multiple edge*** if it appears more than once in $E(G)$. A graph is ***simple*** if it has no loops or multiple edges.

We also say that a graph is *loopless* if multiple edges are allowed but loops are not.

Remark. Unless explicitly stated otherwise, we will only consider simple graphs. General (potentially non-simple) graphs are also called multigraphs.

Bibliography

[Ahl79] L. V. Ahlfors. *Complex Analysis*. McGraw-Hill, 1979.

[Apo57] T. M. Apostol. *Mathematical Analysis*. Addison-Wesley, 1957.

[Axl20] S. Axler. *Measure, Integration & Real Analysis*. Springer, 2020.

[Axl24] S. Axler. *Linear Algebra Done Right*. Springer, 2024.

[Die17] R. Diestel. *Graph Theory*. Springer, 2017.

[DF04] D. S. Dummit and R. M. Foote. *Abstract Algebra*. John Wiley & Sons, 2004.

[Fol99] G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, Inc., 1999.

[HS65] E. Hewitt and K. Stromberg. *Real and Abstract Analysis*. Springer-Verlag, 1965.

[Lan05] S. Lang. *Undergraduate Algebra*. Springer, 2005.

[Mun00] J. R. Munkres. *Topology*. Prentice Hall, Inc., 2000.

[Pól45] G. Pólya. *How to Solve It*. Princeton University Press, 1945.

[Rud76] W. Rudin. *Principles of Mathematical Analysis*. McGraw–Hill, 1976.

[Sch92] A. H. Schoenfeld. "Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics". In: *Handbook for Research on Mathematics Teaching and Learning*. Macmillan, 1992, pp. 334–370.

[Spi65] M. Spivak. *Calculus on Manifolds: A Modern Approach to Classical Theorems of Advanced Calculus*. Harper Collins Publishers, 1965.

Index