Harvard Math 55

Lectures by Denis Auroux, Notes by Gary Hu

Fall 2020 - Spring 2021

From Fall 2020 to Spring 2021, Denis Auroux taught Math 55, a two-semester course divided into two parts:

- Part A: Studies in Algebra and Group Theory
- Part B: Studies in Real and Complex Analysis

Together, they consisted of 73 lectures and 24 problem sets.

This an unofficial set of notes scribed by Gary Hu, who is responsible for all mistakes. If you do find any errors, please report them to: gh7@williams.edu

Contents

1 Group Theory I

1.1 Groups and Their Examples

Groups are abstract algebraic structures that model common features of concrete objects such as numbers, permutations, linear transformations, symmetries, and more.

Definition 1.1. *A **group** G is a set S together with a **law of composition** (a binary operation)*

$$m : S \times S \to S, \quad (a, b) \mapsto a \cdot b,$$

satisfying the following axioms:

1. ***Identity element**: There exists an element $e \in S$ such that, for all $a \in S$,*

$$a \cdot e = e \cdot a = a.$$

2. ***Inverses**: For each $a \in S$, there exists an element $b \in S$ such that*

$$a \cdot b = b \cdot a = e.$$

   *The element b is the **inverse** of a and is often denoted by a^{-1}.*

3. ***Associativity**: For all $a, b, c \in S$,*

$$(a \cdot b) \cdot c = a \cdot (b \cdot c).$$

Remark 1.2. *The following properties follow directly from the group axioms:*

- *The identity element e is unique. If e and e' both satisfy the identity property, then $e = ee' = e'$.*
- *Each element in S has a unique inverse. If b and b' are both inverses of a, then $b = b'$.*
- *The cancellation law holds: for all $a, b, c \in S$, if $a \cdot b = a \cdot c$, then $b = c$. Similarly, if $b \cdot a = c \cdot a$, then $b = c$. This follows from the existence of inverses:*

$$a \cdot b = a \cdot c \implies a^{-1} \cdot (a \cdot b) = a^{-1} \cdot (a \cdot c) \implies b = c.$$

Several structures related to groups arise by relaxing certain axioms:

1. If the axiom of inverses is omitted, the resulting structure is a **semigroup**.
2. A group in which the law of composition is commutative (i.e., $a \cdot b = b \cdot a$ for all $a, b \in S$) is an **abelian group**.

Here are some examples of groups:

Example 1.3 (The Trivial Group)**.** *The set $G = \{e\}$ with the composition rule $e \cdot e = e$. While simple, this is a valid group and serves as a building block in group theory.*

Example 1.4 (Number Systems Under Addition)**.**

$$(\mathbb{Z}, +), \quad (\mathbb{Q}, +), \quad (\mathbb{R}, +), \quad (\mathbb{C}, +),$$

where the identity is 0 and the inverse of x is $-x$. However, $(\mathbb{N}, +)$ (the natural numbers under addition) is not a group, as inverses do not exist.

Example 1.5 (A Group With 2 Elements)**.** *Let $G = \{e, x\}$, where e is the identity and x satisfies $x \cdot x = e$. Examples include:*

- *$\{0, 1\}$ with addition modulo 2, where $1 + 1 = 0$.*
- *$\{+1, -1\}$ with multiplication, where $(-1) \cdot (-1) = +1$.*

Example 1.6 (Cyclic Groups)**.** *The set $\mathbb{Z}/n = \{0, 1, \ldots, n-1\}$ under addition modulo n:*

$$(a + b) \mod n = \begin{cases} a + b & \text{if } a + b < n, \\ a + b - n & \text{otherwise.} \end{cases}$$

This is a finite group of order n. Similarly, the set $[0, 1)$ with addition modulo 1 (denoted $\mathbb{R}/\mathbb{Z}$) is an infinite cyclic group.

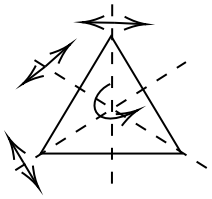
Example 1.7 (Nonzero Numbers Under Multiplication)**.** *The sets*

$$\mathbb{Q}^* = \mathbb{Q} \setminus \{0\}, \quad \mathbb{R}^* = \mathbb{R} \setminus \{0\}, \quad \mathbb{C}^* = \mathbb{C} \setminus \{0\},$$

with the operation of multiplication form groups, where the identity is 1 and the inverse of x is $1/x$. Inside $\mathbb{C}^*$, the set of complex numbers with modulus 1 (denoted S^1) also forms a group under multiplication. These groups are abelian.

Example 1.8 (Symmetries and Permutations)**.**

- *A **permutation** of a set A is a bijection $f : A \to A$. The set of all permutations of A, with composition as the operation, forms a group, denoted $Perm(A)$.*
- *The **symmetric group** on n elements, denoted S_n, is the group of all permutations of the set $\{1, 2, \ldots, n\}$. For example, S_3 has a geometric interpretation as the group of symmetries of an equilateral triangle. These symmetries include three rotations (including the identity) and three reflections.*



Example 1.9 (Matrix Groups)**.**

- *The* ***general linear group****,* $GL_n(\mathbb{R})$*, is the group of all invertible* $n \times n$ *matrices with real entries, under matrix multiplication.*
- *The* ***special linear group****,* $SL_n(\mathbb{R})$*, consists of* $n \times n$ *matrices with real entries and determinant* 1*. Both groups generalize to matrices with coefficients in* $\mathbb{C}$*,* $\mathbb{Q}$*, or* $\mathbb{Z}/n$*.*

1.2 Products of Groups

Definition 1.10. *Let* G *and* H *be groups. The* ***product group*** *of* G *and* H *is the set*

$$G \times H = \{(g, h) \mid g \in G, h \in H\},$$

with the composition law defined by

$$(g, h) \cdot (g', h') = (gg', hh'),$$

where $g, g' \in G$ *and* $h, h' \in H$*.*

Proposition 1.11. *If* G *and* H *are finite groups with orders* $|G| = m$ *and* $|H| = n$*, then the product group* $G \times H$ *is a finite group with order* $|G \times H| = mn$*.*

Remark 1.12. *The result above generalizes to the product of* n *finite groups. Specifically, if* $G_1, G_2, \ldots, G_n$ *are finite groups with orders* $|G_i| = m_i$*, then the product group*

$$G_1 \times G_2 \times \cdots \times G_n$$

is a finite group with order

$$|G_1 \times G_2 \times \cdots \times G_n| = m_1 m_2 \cdots m_n.$$

Definition 1.13. *Let* $\{G_i\}_{i=1}^{\infty}$ *be an infinite collection of groups. We define the following:*

1. *The* ***direct product*** *is given by*

$$\prod_{i=1}^{\infty} G_i = \{(a_1, a_2, \ldots) \mid a_i \in G_i \text{ for all } i\}.$$

2. *The* ***direct sum*** *is the subset of the direct product defined as*

$$\bigoplus_{i=1}^{\infty} G_i = \{(a_1, a_2, \ldots) \mid a_i \in G_i, \text{ all but finitely many } a_i \text{ are the identity element of } G_i\}.$$

Example 1.14. *Let* $G_0 = G_1 = G_2 = \cdots = (\mathbb{R}, +)$*, the additive group of real numbers. Denote an element* $(a_0, a_1, \ldots)$ *by the formal series* $\sum a_i x^i$*. Then:*

1. *The direct product* $\prod_{i=0}^{\infty} \mathbb{R} = \mathbb{R}[\![x]\!]$ *is the set of all formal power series* $\sum_{i=0}^{\infty} a_i x^i$*, with addition defined componentwise.*
2. *The direct sum* $\bigoplus_{i=0}^{\infty} \mathbb{R} = \mathbb{R}[x]$ *is the set of polynomials* $\sum_{finite} a_i x^i$*, where all but finitely many coefficients* a_i *are zero. Addition is again defined componentwise.*

1.3 Subgroups

Definition 1.15. *A* ***subgroup*** H *of a group* G *is a nonempty subset* $H \subset G$ *that satisfies the following conditions:*

- ***Closure under composition:*** *For all* $a, b \in H$*, we have* $ab \in H$*.*
- ***Closure under inversion:*** *For all* $a \in H$*, we have* $a^{-1} \in H$*.*

Since $H \neq \emptyset$, the above conditions imply that the identity element $e \in H$. Thus, H (with the same operation as G) is itself a group.

Remark 1.16. *A subgroup* H *of* G *is* ***proper*** *if* $H \subsetneq G$*.*

Let's take a look at some examples of subgroups:

Example 1.17.

- $(\mathbb{Z}, +) \subset (\mathbb{Q}, +) \subset (\mathbb{R}, +) \subset (\mathbb{C}, +)$
- $(\mathbb{Q}^*, \times) \subset (\mathbb{R}^*, \times) \subset (\mathbb{C}^*, \times) \supset (S^1, \times)$*, where* $S^1 = \{z \in \mathbb{C} \mid |z| = 1\}$*.*
- *The trivial subgroup* $\{e\} \subset G$*.*
- *If* $H_i \subset G_i$ *for* $i = 1, \ldots, n$*, then* $H_1 \times \cdots \times H_n \subset G_1 \times \cdots \times G_n$*.*
- $\bigoplus G_i \subset \prod G_i$*.*

Proposition 1.18. *Let* $a \in \mathbb{Z}_{>0}$*. Then the set*

$$\mathbb{Z}a = \{na \mid n \in \mathbb{Z}\} \subset \mathbb{Z}$$

is a subgroup of $(\mathbb{Z}, +)$*. In fact, every nontrivial subgroup of* $(\mathbb{Z}, +)$ *is of this form.*

Proof. This result follows from the Euclidean algorithm. Let $H \subset \mathbb{Z}$ be a nontrivial subgroup, i.e., $H \neq \{0\}$. Then there exists $a \in H$ such that $a > 0$. Let a_0 be the smallest positive element of H. For any $b \in H$, we can write $b = qa_0 + r$ for some integers $q \in \mathbb{Z}$ and $0 \leq r < a_0$ (remainder). Since $b, qa_0 \in H$, we also have $r \in H$.

By the minimality of a_0, it must be that $r = 0$. Thus, $b \in \mathbb{Z}a_0$, and hence $H \subset \mathbb{Z}a_0$. Conversely, it is clear that $\mathbb{Z}a_0 \subset H$. Therefore, $H = \mathbb{Z}a_0$. □

Thus, every subgroup of $(\mathbb{Z}, +)$ is **generated** by a single element a_0.

Proposition 1.19. *If H and H' are subgroups of a group G, then their intersection $H \cap H'$ is also a subgroup of G.*

Proof. Since $e \in H$ and $e \in H'$, we have $e \in H \cap H'$. Thus, $H \cap H' \neq \emptyset$. For any $a, b \in H \cap H'$, we know that $a, b \in H$ and $a, b \in H'$, so $ab \in H$ and $ab \in H'$. Therefore, $ab \in H \cap H'$. Similarly, $a^{-1} \in H \cap H'$ since inverses are closed in both H and H'. Thus, $H \cap H'$ is a subgroup. □

Remark 1.20. *This result generalizes to intersections of arbitrarily many subgroups.*

Definition 1.21. *Let $S \subset G$ be a nonempty subset of a group G. The **subgroup generated by** S, denoted $\langle S \rangle$, is the smallest subgroup of G containing S. It is given by*

$$\langle S \rangle = \{a_1 \cdots a_k \mid a_i \in S \cup S^{-1},\ 1 \leq i \leq k\}.$$

Definition 1.22. *A group is **cyclic** if it can be generated by a single element.*

Example 1.23. *Examples of cyclic groups include:*

1. *The groups $(\mathbb{Z}, +)$ and $(\mathbb{Z}/n, +)$ are cyclic. In fact, these are the only cyclic groups (up to isomorphism).*
2. *A more advanced example: The group $SL_2(\mathbb{Z})$ can be generated by two elements:*
$$A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Proposition 1.24 (Cayley's Theorem)**.** *Every finite group G is isomorphic to a subgroup of the symmetric group S_n for some n.*

Proof. Define a map $\varphi : G \to \mathrm{Perm}(G)$ by

$$\varphi(g) = m_g,$$

where $m_g : G \to G$ is left multiplication by g, i.e., $m_g(x) = gx$ for all $x \in G$.

To show φ is a homomorphism, note that for $g, h \in G$, we have

$$\varphi(gh)(x) = (gh)x \quad \text{and} \quad (\varphi(g) \circ \varphi(h))(x) = g(hx).$$

By associativity, these are equal, so $\varphi(gh) = \varphi(g) \circ \varphi(h)$.

If $g \neq g'$, then $m_g(e) = g \neq g' = m_{g'}(e)$, so $\varphi(g) \neq \varphi(g')$. Thus, φ is injective. Therefore, $G \simeq \mathrm{Im}(\varphi) \subset \mathrm{Perm}(G) \simeq S_{|G|}$. □

1.4 Homomorphisms

Definition 1.25. *Given two groups G and H, a* ***homomorphism*** $\varphi : G \to H$ *is a map that respects the group operation:*

$$\forall a, b \in G, \quad \varphi(ab) = \varphi(a)\varphi(b).$$

This definition immediately implies:

$$\varphi(e_G) = e_H \quad \text{and} \quad \varphi(a^{-1}) = \varphi(a)^{-1}, \quad \forall a \in G.$$

Remark 1.26. *A pedantic way to state $\varphi(ab) = \varphi(a)\varphi(b)$ is by a commutative diagram:*

$$\begin{array}{ccc} G \times G & \xrightarrow{\varphi \times \varphi} & H \times H \\ \downarrow{\scriptstyle m_G} & & \downarrow{\scriptstyle m_H} \\ G & \xrightarrow[\varphi]{} & H \end{array}$$

Here, the multiplication maps m_G and m_H denote the group operation in G and H, respectively. The diagram is commutative, meaning that the two paths from the top-left corner to the bottom-right corner result in the same map. This formalizes the idea that it does not matter whether we first multiply in G and then apply φ, or apply φ first and then multiply in H.

Example 1.27. *Examples of homomorphisms include:*

- *The modulo map: $\mathbb{Z} \twoheadrightarrow \mathbb{Z}/n$, $a \mapsto a \pmod n$, which sends integers to their remainder modulo n.*
- *If $n \mid m$, the map $\mathbb{Z}/m \twoheadrightarrow \mathbb{Z}/n$ defined similarly, e.g., $\mathbb{Z}/100 \to \mathbb{Z}/10$ maps the last two digits to the last digit.*
- *The determinant map:* $\det : GL_n(\mathbb{R}) \to (\mathbb{R}^*, \times)$, *where* $\det(AB) = \det(A)\det(B)$.

Definition 1.28.

- *A* ***group isomorphism*** *is a bijective homomorphism.*
- *A* ***group automorphism*** *is an isomorphism $\varphi : G \to G$.*

Example 1.29. *Examples of isomorphisms include:*

- *All groups of order 2 are isomorphic: $S_2 = (\{id, (1\,2)\}, \circ) \cong (\{\pm 1\}, \times) = (\mathbb{Z}/2, +)$, since their Cayley tables are identical:*

$\circ$	e	x
e	e	x
x	x	e

- $(\mathbb{R}, +) \xrightarrow{\sim} (\mathbb{R}_+, \times), \quad t \mapsto e^t$.

- $(\mathbb{R}/\mathbb{Z}, +) \xrightarrow{\sim} (S^1, \times), \quad t \mapsto e^{2\pi i t}$.
- *S_3 (the symmetric group on 3 elements) is isomorphic to the group of symmetries of a triangle.*

Definition 1.30. *The **kernel** of a group homomorphism $\varphi : G \to H$ is the set*

$$\ker(\varphi) = \{a \in G \mid \varphi(a) = e_H\}.$$

Proposition 1.31. *The kernel of a homomorphism is a subgroup of G.*

Proof. Check that $\ker(\varphi)$ contains e_G, is closed under the group operation, and is closed under taking inverses. □

Proposition 1.32. *A homomorphism $\varphi : G \to H$ is injective if and only if $\ker(\varphi) = \{e_G\}$.*

Proof. The condition $\varphi(a) = \varphi(b)$ is equivalent to $a^{-1}b \in \ker(\varphi)$. Thus, if $\ker(\varphi) = \{e_G\}$, then φ is injective. □

Definition 1.33. *The **image** of a group homomorphism $\varphi : G \to H$ is the set*

$$Im(\varphi) = \varphi(G) = \{b \in H \mid \exists a \in G \text{ such that } \varphi(a) = b\}.$$

Proposition 1.34. *The image of a homomorphism is a subgroup of H.*

Proposition 1.35. *A homomorphism $\varphi : G \to H$ is surjective if and only if $Im(\varphi) = H$.*

Remark 1.36. *If $\varphi : G \to H$ is injective, then G is isomorphic to the subgroup $Im(\varphi) \subset H$. The isomorphism is given by the map $G \to Im(\varphi)$, $a \mapsto \varphi(a)$.*

Example 1.37. *Let $a \in G$ be any element of a group G. Then the map $\varphi : \mathbb{Z} \to G$, $n \mapsto a^n$, is a homomorphism with image $\langle a \rangle$, the subgroup generated by a.*

Definition 1.38. *The **order** of an element $a \in G$ is the smallest positive integer k such that $a^k = e$, if such k exists. If no such k exists, a has infinite order.*

Proposition 1.39. *If a has infinite order, the powers of a are all distinct, $\varphi : n \mapsto a^n$ is injective, and $\langle a \rangle \cong \mathbb{Z}$. If a has finite order k, then $\ker(\varphi) = k\mathbb{Z}$ and $\langle a \rangle \cong \mathbb{Z}/k$.*

Remark 1.40. *Do not confuse the **order** of an element $a \in G$ with the **order** of the group G, which is the cardinality of G. However, order$(a) = |\langle a \rangle|$.*

Example 1.41. *The group $\mathbb{Z}/6$ is isomorphic to $\mathbb{Z}/2 \times \mathbb{Z}/3$, with the map*

$$a \mapsto (a \pmod 2, a \pmod 3).$$

The element $(1,1) \in \mathbb{Z}/2 \times \mathbb{Z}/3$ *has order 6, so it generates the group. Similarly, if* $\gcd(m,n) = 1$*, then* $\mathbb{Z}/m \times \mathbb{Z}/n \cong \mathbb{Z}/mn$*. However,* $\mathbb{Z}/2 \times \mathbb{Z}/2 \ncong \mathbb{Z}/4$ *because, in* $\mathbb{Z}/2 \times \mathbb{Z}/2$*,* $x + x = 0$ *for all* x*, whereas in* $\mathbb{Z}/4$*,* $1 + 1 \neq 0$*.*

1.5 Interlude: Set Theory

Definition 1.42. *A map of sets* $f : S \to T$ *is:*

- ***Injective*** *if* $\forall a, b \in S$*,* $f(a) = f(b) \implies a = b$ *(equivalently,* $a \neq b \implies f(a) \neq f(b)$*). We denote this as* $f : S \hookrightarrow T$*.*
- ***Surjective*** *if* $\forall c \in T, \exists a \in S$ *such that* $f(a) = c$*. We denote this as* $f : S \twoheadrightarrow T$*.*
- ***Bijective*** *if* f *is both injective and surjective.*

Definition 1.43. *Two sets* S *and* T *have the* ***same cardinality*** *if there exists a bijection* $f : S \to T$*. In this case, we write* $|S| = |T|$*. If there exists an injection* $f : S \hookrightarrow T$*, we write* $|S| \leq |T|$*.*

This notation works because of the following theorem.

Proposition 1.44 (Schröder-Bernstein Theorem)**.** *If there exist injective maps* $f : S \hookrightarrow T$ *and* $g : T \hookrightarrow S$*, then* $|S| = |T|$*.*

Proof. See Halmos' *Naive Set Theory*, page 88, for a complete proof. The proof constructs a bijection $S \xrightarrow{\sim} T$ by using f on a subset of S and g^{-1} on the complement. □

Example 1.45. *The sets* $\mathbb{N}$*,* $\mathbb{Z}$*, and* $\mathbb{Q}$ *all have the same cardinality and are* ***countably infinite****.*

To define a bijection $\mathbb{N} \to \mathbb{Z}$*, consider the piecewise function:*

$$f(n) = \begin{cases} \frac{n}{2} & \text{if } n \text{ is even,} \\ -\frac{n+1}{2} & \text{if } n \text{ is odd.} \end{cases}$$

For $\mathbb{Q}$*, we eed to enumerate* $\mathbb{N} \times \mathbb{N}$ *to establish a bijection between* $\mathbb{N}$ *and* $\mathbb{Q}$*.*

Example 1.46. *In contrast, the set* $\mathbb{R}$ *is* ***uncountable****, as demonstrated by Cantor's diagonal argument. No map* $f : \mathbb{N} \to \mathbb{R}$ *can be surjective. To see this, consider the decimal or binary expansion of the numbers in the image of* f*:*

$$\begin{aligned} f(0) &= a_{00}a_{01}a_{02}a_{03}\ldots \\ f(1) &= a_{10}a_{11}a_{12}a_{13}\ldots \\ f(2) &= a_{20}a_{21}a_{22}a_{23}\ldots \\ f(3) &= a_{30}a_{31}a_{32}a_{33}\ldots \end{aligned}$$

Now define a number $y = b_0b_1b_2b_3\ldots$ *where* $b_i \neq a_{ii}$ *for each* i*. By construction,* $y \neq f(i)$ *for all* $i \in \mathbb{N}$*, so* f *cannot be surjective.*

The same argument generalizes to show that there are arbitrarily large cardinalities. For any set S, consider its ***power set*** $\mathcal{P}(S) = \{\text{subsets of } S\}$*, which satisfies* $|\mathcal{P}(S)| > |S|$*. We can represent the power set as* $\{0,1\}^S$*, the set of maps* $f : S \to \{0,1\}$*. This isomorphism is defined as follows:*

$$A \mapsto (\mathbb{1}_A : x \mapsto \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases})$$

Conversely, any map $f : S \to \{0,1\}$ *corresponds to the subset* $f^{-1}(1) \subset S$*.*

If S *is finite with* $|S| = n$*, then* $|\mathcal{P}(S)| = 2^n$*.*

What happens if S is infinite?

Proposition 1.47. *If* S *is infinite, then* $|\mathcal{P}(S)| > |S|$*.*

Proof. Given any map $f : S \to \mathcal{P}(S)$, define the set

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

Assume, for contradiction, that $A = f(a)$ for some $a \in S$. Then:

$$a \in A \iff a \notin f(a) = A,$$

a contradiction. Thus, $A \notin f(S)$, and f cannot be surjective. □

1.6 Classification of Finite Groups

Proposition 1.48. *Every finite group* G *is isomorphic to a subgroup of the symmetric group* S_n *for some* n*.*

In fact, we can take $n = |G|$. While this result establishes that every finite group embeds into a symmetric group, it is not particularly useful for classifying finite groups, as subgroups of S_n are notoriously challenging to classify in general.

Proof. Define a map $\phi : G \to \mathrm{Perm}(G)$ by $\phi(g) = m_g$, where m_g is the left multiplication by g given by:

$$m_g : G \to G, \quad x \mapsto gx.$$

To see that m_g is a permutation, note that it is a bijection since group multiplication is invertible. The map ϕ is a homomorphism because:

$$\phi(gh) = m_{gh}, \quad m_{gh}(x) = (gh)x = g(hx),$$

which is the same as $\phi(g) \circ \phi(h) = m_g \circ m_h$.

To show injectivity, assume $g \neq g'$. Then:

$$m_g(e) = g \neq g' = m_{g'}(e),$$

so $\phi(g) \neq \phi(g')$. Hence, ϕ is injective, and we have $G \simeq \mathrm{Im}(\phi) \subset \mathrm{Perm}(G) \simeq S_{|G|}$. □

An important question in group theory is the classification of finite groups up to isomorphism. This problem becomes increasingly challenging as $|G|$ grows. Here are some foundational results for small group orders:

- Every group of order 2 is isomorphic to $\mathbb{Z}/2$. This can be verified by writing out the group operation table.
- Similarly, every group of order 3 is isomorphic to $\mathbb{Z}/3$.
- For groups of order 4, there are exactly two isomorphism classes: $\mathbb{Z}/4$ and $\mathbb{Z}/2 \times \mathbb{Z}/2$. These groups are distinct:
  - In $\mathbb{Z}/2 \times \mathbb{Z}/2$, every nonzero element has order 2.
  - In $\mathbb{Z}/4$, there exists an element of order 4.

  Hence, these are the only two groups of order 4 up to isomorphism.

The full classification of finite groups was completed in the 1980s and spans thousands of pages of mathematical work. While we will explore some of the key tools and concepts in this course, the complete classification is far beyond our scope.

1.7 Interlude: Equivalence Relations and Partitions

An **equivalence relation** on a set S is a way to declare certain elements equivalent to each other (denoted "$a \sim b$"), resulting in a smaller set of equivalence classes "$S/\sim$" (the quotient of S by $\sim$).

Definition 1.49. *An **equivalence relation** on a set S is a binary relation (i.e., a map $\sim\colon S \times S \to \{0,1\}$, or equivalently a subset of $S \times S$; we write $a \sim b$ if and only if (a,b) is in this subset) which satisfies the following properties:*

1. ***Reflexive**: $\forall a \in S, a \sim a$.*
2. ***Symmetric**: $\forall a, b \in S, a \sim b \implies b \sim a$.*
3. ***Transitive**: $\forall a, b, c \in S$, if $a \sim b$ and $b \sim c$, then $a \sim c$.*

*The **equivalence class** of $a \in S$ is defined as:*

$$[a] = \{a' \in S \mid a \sim a'\}.$$

By the transitivity property, all elements of $[a]$ are equivalent to each other.

Proposition 1.50. *The equivalence classes form a **partition** of S. That is, the equivalence classes are mutually disjoint subsets of S, and their union equals S.*

Definition 1.51. *The **quotient** of S by $\sim$ is the set of **equivalence classes**:*

$$S/\sim = \{[a] \mid a \in S\} \subset \mathcal{P}(S),$$

where $\mathcal{P}(S)$ is the power set of S. This comes with a natural surjective map:

$$S \to S/\sim, \quad a \mapsto [a].$$

Example 1.52.

1. *Let $S = \mathbb{Z}$. For a fixed $n \in \mathbb{Z}_{>0}$, define $a \sim b$ if and only if $n \mid (b-a)$. This is the congruence relation modulo n, which can be verified to be an equivalence relation. The equivalence classes are:*

   $$[0] = \{\ldots, -n, 0, n, 2n, \ldots\} = n\mathbb{Z}, \quad [1] = \{\ldots, 1-n, 1, 1+n, 1+2n, \ldots\},$$

   continuing up to $[n-1]$. There are n distinct equivalence classes. The quotient is naturally in bijection with $\mathbb{Z}/n$:

   $$\mathbb{Z} \twoheadrightarrow \mathbb{Z}/\sim \simeq \mathbb{Z}/n, \quad a \mapsto [a].$$

   Although $\mathbb{Z}/n$ is often written as $\{0, \ldots, n-1\}$ to avoid the language of equivalence classes, it is more accurate to redefine it as the quotient set.

2. *Given a map $f : S \to T$, define $a \sim b$ if and only if $f(a) = f(b)$. This is an equivalence relation, and the partition into equivalence classes is:*

   $$S = \bigsqcup_{t \in T} f^{-1}(t) = \{a \in S \mid f(a) = t\},$$

   where the disjoint union is taken over $t \in f(S) \subset T$. The map f naturally factors through the quotient:

   $$S \twoheadrightarrow S/\sim \hookrightarrow T, \quad a \mapsto [a] \mapsto f(a).$$

   If f is surjective, then $S/\sim \simeq T$.

Using this conclusion, we observe the following equivalence:

$$\begin{aligned} \textit{Equivalence relation on } S &\iff \textit{Partition of } S \textit{ into disjoint subsets} \\ &\iff \textit{Surjective map } S \to T \end{aligned}$$

(up to composition with a bijection $T \xrightarrow{\sim} T'$).

1.8 Cosets and Normal Subgroups

Let $\varphi : G \to H$ be a surjective group homomorphism. Recall that the **kernel** of φ, denoted $K = \mathrm{Ker}(\varphi) = \{a \in G \mid \varphi(a) = e_H\}$, is a subgroup of G. Consider the partition of G induced by φ. We have the following equivalence:

$$\begin{aligned} \varphi(a) = \varphi(b) &\iff \varphi(a)^{-1}\varphi(b) = e_H \\ &\iff a^{-1}b \in K \\ &\iff b \in aK = \{ak \mid k \in K\}. \end{aligned}$$

Definition 1.53. *Given a subgroup K of a group G, the set*

$$aK = \{ak \mid k \in K\} \subset G$$

*is the **left coset** of K in G containing a.*

Proposition 1.54.

- *The relation $a \sim b \iff a^{-1}b \in K$ is an equivalence relation on G, and the equivalence classes are the left cosets of K.*
- *The quotient, denoted by G/K, is the set of left cosets. Therefore, we have a partition $G = \bigsqcup_{aK \in G/K} aK$.*

Proof.

- Reflexivity: For any $a \in G$, we have $a^{-1}a = e \in K$, so $a \sim a$.
- Symmetry: If $a \sim b$, then $a^{-1}b \in K$. Therefore, $(a^{-1}b)^{-1} = b^{-1}a \in K$, so $b \sim a$.
- Transitivity: If $a \sim b$ and $b \sim c$, then $a^{-1}b \in K$ and $b^{-1}c \in K$. Hence, $(a^{-1}b)(b^{-1}c) \in K$, so $a \sim c$.

Additionally, we can check that $b \in aK \iff \exists k \in K$ such that $b = ak \iff \exists k \in K$ such that $a^{-1}b = k \iff a^{-1}b \in K \iff a \sim b$. □

Example 1.55. *Consider the surjective homomorphism $\varphi : \mathbb{Z} \twoheadrightarrow \mathbb{Z}/n, a \mapsto a \pmod n$. The kernel of φ is $\mathbb{Z}n \subset \mathbb{Z}$. The cosets of $\mathbb{Z}n$ are $[k] = k + \mathbb{Z}n$ for $0 \leq k \leq n-1$. We have a bijection $\mathbb{Z}/\mathbb{Z}n \simeq \mathbb{Z}/n$, where $[k] \mapsto k$. This gives rise to a group law on the quotient: coset addition corresponds to addition modulo n.*

When a subgroup K is the kernel of a homomorphism $\varphi : G \twoheadrightarrow H$, we obtain a bijection $G/K \simeq H$ with $aK \mapsto \varphi(a)$. This bijection provides a group structure on G/K, where $(aK) \cdot (bK) = abK$. Thus, the map $G \twoheadrightarrow G/K, a \mapsto aK$ is a group homomorphism.

However, this does not necessarily work for all subgroups $K \subset G$. For example, it fails for $\{e, h\} \subset D_4$.

Analogous to left cosets, we can also define right cosets:

Definition 1.56. *The set*

$$Ka = \{ka \mid k \in K\} \subset G$$

*is the **right coset** of K in G containing a, and it corresponds to the equivalence relation $a \sim b \iff ba^{-1} \in K$.*

Remark 1.57. *Neither left cosets nor right cosets are subgroups of G (except for K itself). However, the set $aKa^{-1} = \{aka^{-1} \mid k \in K\}$ is a subgroup.*

Definition 1.58. *A subgroup $K \subset G$ is a **normal subgroup** if the left cosets equal the right cosets, i.e., $aK = Ka$ for all $a \in G$.*

In other words, the two equivalence relations $a \sim b \iff a^{-1}b \in K$ and $a \sim b \iff ba^{-1} \in K$ must coincide.

Proposition 1.59. *Given a group G and a subgroup $K \subset G$, there exists a group homomorphism $\varphi : G \to H$ with $Ker(\varphi) = K$ if and only if K is a normal subgroup. In this case, G/K inherits a group structure defined by $(aK) \cdot (bK) = abK$, and the map $G \twoheadrightarrow G/K$ is a group homomorphism.*

Proof. **($\implies$ direction):** Suppose there exists a homomorphism $\varphi : G \to H$ with $\mathrm{Ker}(\varphi) = K$. For any $a, b \in G$, we have $\varphi(a) = \varphi(b) \iff \varphi(a)^{-1}\varphi(b) = e \iff \varphi(a^{-1}b) = e \iff a^{-1}b \in K$, which implies $b \in aK$. Similarly, $b \in Ka$. Thus, $aK = Ka$ for all $a \in G$, so K is normal.

($\impliedby$ direction): Assume K is normal. Define an operation on G/K by $(aK) \cdot (bK) = abK$. We need to verify that this operation is well-defined: if $aK = a'K$ and $bK = b'K$, then $a^{-1}a' \in K$ and $b^{-1}b' \in K$. It follows that

$$(ab)^{-1}(a'b') = b^{-1}a^{-1}a'b' \in K,$$

using the normality of K. This operation satisfies the group axioms. Furthermore, the map $G \twoheadrightarrow G/K, a \mapsto aK$ is a well-defined group homomorphism with kernel K. □

Example 1.60.

- *Any subgroup of an abelian group is normal.*
- *In D_4, the subgroup $\{e, h\}$ is not normal. However, the subgroup generated by the horizontal and vertical reflections is normal, and the quotient is isomorphic to $\mathbb{Z}/2$.*
- *In any group G, the **center** $Z(G) = \{z \in G \mid az = za \text{ for all } a \in G\}$ is a normal subgroup. To verify that $Z(G)$ is indeed a subgroup, observe that for any $a \in G$ and $z \in Z(G)$, we have $a^{-1}za = z$ for all $z \in Z(G)$. This demonstrates that $Z(G)$ is invariant under conjugation, which is stronger than normality, where we only require $a^{-1}za$ to lie in $Z(G)$ (not necessarily equal to z).*

Earlier, we discussed the partition of a group G into (left) cosets of a subgroup $H \subset G$, where $aH = \{ah \mid h \in H\} \subset G$.

Definition 1.61. *The **cosets** of H in G are the equivalence classes under the relation $a \sim b \iff a^{-1}b \in H$. The quotient G/H is the set of cosets, and the **index** of the subgroup H in G is the number of cosets, denoted by $(G : H) = |G/H|$.*

When G is a finite group, each coset has cardinality $|aH| = |H|$, since the map $a \mapsto aH$, with $h \mapsto ah$, is a bijection. This implies that the partition $G = \bigsqcup_{aH \in G/H} aH$ leads to Lagrange's Theorem:

Theorem 1.62 (Lagrange's Theorem)**.** *If H is a subgroup of a finite group G, then*

$$|G| = |G/H| \times |H|.$$

Corollary 1.63. *If H is a subgroup of a finite group G, then $|H|$ divides $|G|$.*

Corollary 1.64. *For any element $a \in G$ in a finite group, the order of a divides $|G|$.*

Corollary 1.65. *If $|G| = p$ is prime, then $G \simeq \mathbb{Z}/p$.*

Proof. Take any $a \in G$ such that $a \neq e$. The order of a is p, so $\langle a \rangle = G$, and $G = \{e, a, \ldots, a^{p-1}\}$. We define a bijection $G \xrightarrow{\sim} \mathbb{Z}/p$ by mapping $a^k \mapsto k \pmod{p}$. □

Example 1.66. *Consider S_3, the group of permutations of $\{1, 2, 3\}$. We have:*

- *e, the identity element, which does nothing, has order 1.*
- *Three transpositions that swap two elements: $(1\,2)$, $(2\,3)$, and $(1\,3)$, which are the reflections of the triangle; each has order 2.*
- *Two 3-cycles: $(1\,2\,3)$ and $(1\,3\,2)$, corresponding to rotations by $\pm 120°$; each has order 3.*

The subgroups of S_3 have orders 1, 2, 3, or 6, and any subgroup of order 2 or 3 is necessarily cyclic:

- *$\{e\}$ is trivial.*
- *$\{e, (1\,2)\}$ and two others are isomorphic to $\mathbb{Z}/2$.*
- *$\{e, (1\,2\,3), (1\,3\,2)\}$ is a subgroup of rotations, isomorphic to $\mathbb{Z}/3$.*
- *The entire group S_3.*

Which ones are normal subgroups?

1. *$\{e\}$ and S_3 are obviously normal subgroups.*
2. *$\{e, (1\,2)\}$ is not normal because its conjugate, $(1\,2\,3)\{e, (1\,2)\}(1\,2\,3)^{-1} = \{e, (2\,3)\}$, is not equal to $\{e, (1\,2)\}$.*
3. *$\{e, (1\,2\,3), (1\,3\,2)\} \simeq \mathbb{Z}/3$ is normal: It's the kernel of the homomorphism $S_3 \overset{sign}{\twoheadrightarrow} \{\pm 1\} \simeq \mathbb{Z}/2$, where rotations map to $+1$ and reflections map to -1, corresponding to the determinant of the corresponding 2×2 matrix.*

Definition 1.67. *A group G is **simple** if it has no normal subgroups other than G and $\{e\}$.*

1.9 Exact Sequences

Definition 1.68. *A sequence of groups and homomorphisms*

$$\cdots \to G_{i-1} \overset{\varphi_{i-1}}{\to} G_i \overset{\varphi_i}{\to} G_{i+1} \to \dots$$

*is an **exact sequence** if, for all i, the image of φ_{i-1} equals the kernel of φ_i, i.e.,*

$$Im(\varphi_{i-1}) = Ker(\varphi_i).$$

This condition means that for each $x \in G_i$, $\varphi_i(x) = e \iff \exists a \in G_{i-1}$ such that $x = \varphi_{i-1}(a)$. In particular, we have $\varphi_i \circ \varphi_{i-1} = 0$, where 0 denotes the trivial homomorphism.

Definition 1.69. *A **short exact sequence** is a special case of an exact sequence, which has the form*

$$\{e\} \to A \overset{\varphi}{\to} G_i \overset{\psi}{\to} G_{i+1} \to \{e\},$$

where φ is an injective homomorphism, ψ is a surjective homomorphism, and $Im(\varphi) = Ker(\psi)$.

Proposition 1.70. *A short exact sequence of the form*

$$\{e\} \to A \overset{\varphi}{\to} G_i \overset{\psi}{\to} G_{i+1} \to \{e\}$$

exists if and only if there is a normal subgroup $K \cong A$ of B, such that the quotient group $B/K \cong C$.

The prototypical short exact sequence is given by

$$1 \to K \overset{\text{inclusion}}{\to} B \overset{\text{quotient}}{\to} B/K \to 1.$$

Example 1.71.

1. *For any groups A and C, the following sequence is exact:*

$$\{e\} \to A \to A \times C \to C \to \{e\}, \quad a \mapsto (a, e),\ (a, c) \mapsto c.$$

2. *The sequence*

$$0 \to \mathbb{Z}/2 \to \mathbb{Z}/6 \to \mathbb{Z}/3 \to 0, \quad n \mapsto 3n,\ m \mapsto m \pmod 3$$

is exact, as well as the sequence

$$0 \to \mathbb{Z}/3 \to \mathbb{Z}/6 \to \mathbb{Z}/2 \to 0, \quad n \mapsto 2n,\ m \mapsto m \pmod 2.$$

3. *There exists an exact sequence*

$$\{e\} \to \mathbb{Z}/3 \to S_3 \overset{sign}{\to} \mathbb{Z}/2 \to \{e\}, \quad n \mapsto (123)^n,$$

but there is no exact sequence

$$\{e\} \to \mathbb{Z}/2 \to S_3 \overset{sign}{\to} \mathbb{Z}/3 \to \{e\}$$

because there is no normal subgroup of S_3 of order 2.

1.10 More About Symmetric Groups

Definition 1.72. *A **cycle** $\sigma = \{a_1\ a_2\ \dots\ a_k\} \in S_n$, where each a_i is a distinct element of $\{1, \dots, n\}$, is a permutation that maps $a_1 \mapsto a_2$, $a_2 \mapsto a_3$, $\dots$, $a_k \mapsto a_1$, and all other elements are mapped to themselves.*

Proposition 1.73. *Any permutation can be expressed as a product of disjoint cycles, uniquely up to the reordering of the factors.*

Since disjoint cycles commute, the order of multiplication does not matter.

Example 1.74.

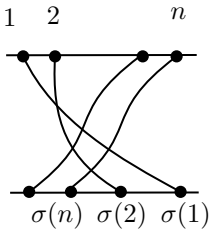
$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 3 & 5 & 6 & 4 & 2 & 1 \end{pmatrix} = (136)(25)$$

Proposition 1.75. *A k-cycle can be written as a product of $(k-1)$ transpositions:*

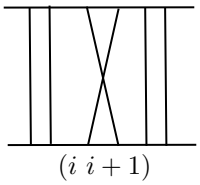
$$(a_1\ a_2\ \dots\ a_k) = (a_1\ a_2) \circ (a_2\ a_3) \circ \cdots \circ (a_{k-1}\ a_k)$$

Therefore, S_n is generated by transpositions $(i\,j)$ where $1 \leq i < j \leq n$. In fact, it is generated by the set of transpositions $(1\,2), (2\,3), \dots, (n-1\,n)$.

The idea for theorems of this kind are as follows: draw σ as



and then slice it into stacks of



Remark 1.76. *Take a look at the bubble sort algorithm.*

Permutations are classified as odd or even depending on the length of the expression of σ as a product of transpositions. This is equivalent to the parity of the set $\#\{(i,j) \mid 1 \leq i < j \leq n, \sigma(j) > \sigma(i)\}$. This fact is nontrivial, and its proof can be done by induction.

The even permutations form a normal subgroup, A_n = alternating subgroup $\subset S_n$, with the exact sequence:

$$1 \to A_n \to S_n \to \mathbb{Z}/2 \to 1$$

Let's see a cool fact. Although $A_3 \simeq \mathbb{Z}/3$ and A_4 has a normal subgroup $\mathbb{Z}/2 \times \mathbb{Z}/2$:

Proposition 1.77. *A_n is simple for $n \geq 5$*

This result is crucial for proving that there is no general formula for solving polynomial equations of degree $n \geq 5$. For instance, the quadratic formula has a $\pm\sqrt{\cdot}$ term, and the sign comes from the fact that there is no consistent choice of square root in $\mathbb{C}$—this ambiguity arises in $\mathbb{Z}/2 \simeq S_2$, which permutes the two roots. The Cardano formula for cubics contains nested roots such as $\sqrt[3]{\cdots + \sqrt{\cdots}}$, and similar ambiguities involving $\mathbb{Z}/2$ and $\mathbb{Z}/3$ combine to form an S_3 group that permutes the roots. Similarly, the formula for the roots of a degree 5 equation must incorporate an S_5 symmetry. However, expressions involving radicals like $\sqrt[k]{\cdots}$ can only involve cyclic groups $\mathbb{Z}/k$, which cannot be isomorphic to S_5, as A_5 is simple.

Here's another cool fact:

Proposition 1.78. *$Aut(S_n) \simeq S_n$ for all n except when $n = 2$ ($Aut(S_2) = \{id\}$) and $n = 6$ ($Aut(S_6) \subsetneq S_6$).*

1.11 Free Groups

We've discussed the **center** $Z(G) = \{z \in G \mid az = za \text{ for all } a \in G\}$. Since elements of the center commute with every element of the group, they also commute with each other, implying that $Z(G)$ is abelian. Additionally, we have $aZ(G)a^{-1} = Z(G)$, so $Z(G)$ is a normal subgroup of G.

Another interesting object is the commutator subgroup:

Definition 1.79. *The **commutator subgroup** is defined as*

$$C(G) = [G, G] = \left\{ \prod_{i=1}^{k} [a_i, b_i] \mid a_i, b_i \in G \right\}$$

where $[a, b] := aba^{-1}b^{-1}$.

Note that the commutator $[a, b] = e$ if and only if $ab = ba$.

Proposition 1.80. *The commutator subgroup is a normal subgroup of G.*

Proof. We compute the conjugate of the commutator subgroup:

$$g^{-1} \prod_{i=1}^{k} [a_i, b_i] g = \prod_{i=1}^{k} [g^{-1}a_i g, g^{-1}b_i g] \implies g^{-1}C(G)g = C(G) \text{ for all } g \in G.$$

□

Definition 1.81. *The quotient $G/[G,G]$ is the **abelianization** of G.*

Since $[G,G]$ contains all commutators $[a,b]$, taking the quotient makes all commutators trivial, i.e., $[a,b] = e$ in the quotient group, which implies that all elements commute with each other in $G/[G,G]$. Because $[G,G]$ is generated by commutators, it is the smallest subgroup of G where this property holds. Therefore, the abelianization is the largest abelian group onto which G admits a surjective homomorphism.

Definition 1.82. *The **free group** F_n on n generators $a_1, \ldots, a_n$ is the collection of all reduced words $a_{i_1}^{n_1} a_{i_2}^{n_2} \ldots a_{i_k}^{n_k}$ of any length $k \geq 0$, where $i_1, \ldots, i_k \in \{1, \ldots, n\}$, $i_j \neq i_{j+1}$, and $n_1, \ldots, n_k$ are non-zero integers. The law of composition is given by juxtaposition, and non-reduced words (where $i_j = i_{j+1}$ for some j or some n_j is zero) are simplified to reduced ones by combining repeated terms and eliminating unnecessary ones. The identity element is represented by the empty word of length $k = 0$.*

This is the "largest" group with n generators, and all other groups with n generators are isomorphic to quotients of F_n. If G is generated by $g_1, \ldots, g_m \in G$, a homomorphism $F_n \to G$ is defined by mapping the word $\prod_{a_{ij}}^{m_j}$ to $\prod_{g_{ij}^{m_j}}$.

Definition 1.83. *A finitely generated group is **finitely presented** if the kernel of the homomorphism $\prod_{a_{ij}}^{m_j} \mapsto \prod_{g_{ij}^{m_j}}$ is the smallest normal subgroup of F_n containing a finite subset $\{r_1, \ldots, r_k\} \subset F_n$ (i.e., the subgroup generated by r_j's and their conjugates $x^{-1} r_j x$).*

If we write $G \simeq \langle a_1, \ldots, a_n \mid r_1, \ldots, r_k \rangle$, where $a_1, \ldots, a_n$ are the generators and $r_1, \ldots, r_k$ are the relations, then $G \simeq F_n / \langle \text{conjugates of } r_1, \ldots, r_k \rangle$.

Example 1.84.

$$\mathbb{Z}^n \simeq \langle a_1, \ldots, a_n \mid a_i a_j a_i^{-1} a_j^{-1} \text{ for all } i, j \rangle$$

Example 1.85.

$$S_3 \simeq \langle t_1, t_2 \mid t_1^2, t_2^2, (t_1 t_2)^3 \rangle$$

2 Linear Algebra I

2.1 Rings and Fields

Now, we proceed to the study of rings and fields on our way to vector spaces.

Definition 2.1. *A **(commutative) ring** is a set R with two operations $+$ and $\times$, such that*

1. *$(R, +)$ is an abelian group with identity $0 \in R$.*
2. *$(R, \times)$ is a commutative semigroup with identity $1 \in R$, i.e., $1a = a1 = a$ for all $a \in R$, and $a(bc) = (ab)c$ for all $a, b, c \in R$. Additionally, $ab = ba$ for all $a, b \in R$ if the operation is commutative.*
3. *The distributive law holds: $a(b + c) = ab + ac$ for all $a, b, c \in R$.*

Definition 2.2. *A **field** K is a commutative ring such that for all $a \neq 0$, there exists $b = a^{-1}$ such that $ab = 1$.*

For example, $(K - \{0\}, \times)$ is an abelian group rather than a semigroup.

Remark 2.3.

- *The ring axioms imply $0a = a0 = 0$ for all $a \in R$: $a0 = a(0+0) = a0+a0$.*
- *The trivial ring $R = \{0\}$ is the only case where $0 = 1$. By convention, this is not considered a field.*
- *Most rings of interest to us are commutative, with matrices being the main exception.*
- *In a field, $ab = 0$ implies $a = 0$ or $b = 0$, but this is not necessarily true in a ring.*
- *In a field, we have the usual cancellation properties for both addition and multiplication.*

Definition 2.4. *A **ring/field homomorphism** is a map $\varphi : R \to S$ that respects both operations:*

$$\varphi(a + b) = \varphi(a) + \varphi(b), \quad \varphi(ab) = \varphi(a)\varphi(b), \quad \varphi(1_R) = 1_S.$$

Thus, $\varphi(0) = 0$ and $\varphi(-a) = -\varphi(a)$. The latter does not follow directly from $\varphi(ab) = \varphi(a)\varphi(b)$, even for fields; for instance, consider the homomorphism $\varphi \equiv 0$.

Proposition 2.5. *If $\varphi : R \to S$ is a field homomorphism, then φ is injective.*

Proof. If $a \neq 0$, then there exists b such that $ab = 1_R$, so $\varphi(a)\varphi(b) = \varphi(ab) = 1_S \neq 0_S$, which implies $\varphi(a) \neq 0_R$. Therefore, $\mathrm{Ker}(\varphi) = \{0\}$, so φ is injective. □

Example 2.6.

- $\mathbb{Z}$, $\mathbb{Z}/n$ *are rings.*
- $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ *are fields. Similarly,* $\mathbb{Z}/p$ *for primes* p*, denoted* $\mathbb{F}_p$*, is a field. This is because, for any non-zero* $k \in \mathbb{Z}/p$*, its order is* p*, so* $\{0, k, 2k, \ldots, (p-1)k\} = \mathbb{Z}/p$*, and there exists* $k \in \{0, \ldots, p-1\}$ *such that* $lk = 1 \pmod{p}$*, providing an inverse.*

Definition 2.7. *Given a field* k*, the* ***ring of polynomials*** *in one formal variable* x *is defined as*

$$k[x] := \{a_0 + a_1 x + \cdots + a_n x^n \mid a_i \in k, n \in \mathbb{N}\}.$$

Remark 2.8. *In the above definition,* x *is a formal variable, i.e., not an element of any set, though we can evaluate a polynomial at any element of a field containing* k*. Thus, a polynomial corresponds to a finite tuple* $(a_0, a_1, \ldots, a_n, 0, 0, \ldots)$ *of elements from* k*, with component-wise addition (but not component-wise multiplication).*

Example 2.9. $k[x]$ *is not a field, but it can be turned into one by considering fractions. The* ***field of rational functions*** *is*

$$k(x) = \left\{ \frac{p}{q} \mid p, q \in k[x], q \neq 0 \right\},$$

where $\frac{p}{q} \sim \frac{p'}{q'}$ *if* $pq' = p'q$*.*

Remark 2.10. *This generalizes to polynomials of rational functions in any number of variables.*

Definition 2.11. *The ring of* ***formal power series*** *in* x *is*

$$K[\![x]\!] = \left\{ \sum_{i=0}^{\infty} a_i x^i \mid a_i \in K \right\}.$$

Addition and multiplication in this ring follow the same rules as for polynomials, performed term by term. It is important to check that each coefficient in $\left(\sum a_i x^i\right)\left(\sum b_j x^j\right)$ is a finite expression.

Lemma 2.12. $\sum a_i x^i$ *has a multiplicative inverse in* $K[\![x]\!]$ *if and only if* $a_0 \neq 0$*.*

Proof. We seek $\sum_{i \geq 0} b_i x^i$ such that

$$\left(\sum_{i \geq 0} a_i x^i \right) \left(\sum_{i \geq 0} b_i x^i \right) = 1.$$

This results in the following system of equations:

$$\begin{aligned} a_0b_0 &= 1, \\ a_0b_1 + a_1b_0 &= 0, \\ a_0b_2 + a_1b_1 + a_2b_0 &= 0, \\ &\dots \end{aligned}$$

If $a_0 = 0$, there is clearly no solution. If $a_0 \neq 0$, we can solve inductively: $b_0 = \frac{1}{a_0}$, $b_1 = -\frac{a_1b_0}{a_0}$, and so on. (Each step involves solving for $b_i = -\frac{\dots}{a_0}$).

Since every nonzero element of $K[\![x]\!]$ can be written as $a_mx^m + a_{m+1}x^{m+1} + \cdots = x^m(a_m + a_{m+1} + \dots)$, where a_m is the first non-zero coefficient and $a_m + a_{m+1} + \dots$ is invertible, to obtain a field, we must allow for x^{-m}. □

Definition 2.13. *The field of* ***Laurent series*** *is*

$$k((x)) = \left\{ \sum_{i=m}^{\infty} a_i x^i \mid m \in \mathbb{Z}, a_i \in K \right\}.$$

Given a field k and a polynomial $f \in k[x]$ of degree ≥ 0, we can evaluate $f(r)$ for $r \in k$ and search for roots $r \in k$ such that $f(r) = 0$. If no roots exist in k, we can form a field $K \supset k$ in which f has a root.

Example 2.14.

1. *For $k = \mathbb{Q}$, the polynomial $x^2 - 2$ has no roots, but we can form the field $\mathbb{Q}(\sqrt{2}) := \{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\}$, which is a field.*
2. *For $k = \mathbb{R}$, the polynomial $x^2 + 1$ leads to the field $\mathbb{R}(\sqrt{-1}) = \mathbb{C}$.*

Given a field k, we always have a ring homomorphism $\varphi : \mathbb{R} \to k$, $1 \mapsto 1_k$. For some fields, this homomorphism is injective. If it is, we say that k has **characteristic zero**.

Proposition 2.15. *$ker(\varphi : \mathbb{Z} \to k) = \mathbb{Z}_p$ for some prime p.*

Proof. $\ker(\varphi)$ is a subgroup of $\mathbb{Z}$, hence must be of the form $\mathbb{Z}n$. If n is not prime, we can write $n = ab$ for some integers a, b such that $1 < a, b < n$. Then, $\varphi(n) = \varphi(ab) = \varphi(a)\varphi(b) = 0 \in k$, but this implies $\varphi(a) = 0$ or $\varphi(b) = 0$. Since n is the smallest positive integer such that $\varphi(n) = 0$, this leads to a contradiction. □

We say that k has **characteristic** p if $\ker(\varphi) = \mathbb{Z}p$. This means $p \cdot 1_k = 1 + \cdots + 1$ (with p summands) equals 0.

Thus far, our only example of such a field is $\mathbb{Z}/p$, though there are more.

Proposition 2.16. *For all $n \geq 1$ and prime p, there exists a unique field with p^n elements (up to isomorphism), and these are all the finite fields.*

There are also infinite fields of characteristic p, such as $\mathbb{Z}/p((x))$.

2.2 Vector Spaces

Definition 2.17. *Fix a field k. A **vector space** over k is a set V with two operations:*

1. *Addition $+$: $V \times V \to V$ such that $(V, +)$ is an abelian group with identity $0 \in V$.*
2. *Scalar multiplication $\cdot$: $k \times V \to V$ which is associative: $(ab)v = a(bv)$, $1v = v$, $0v = 0$, and distributive: $a(v + v') = av + av'$, $(a + b)v = av + bv$.*

Definition 2.18. *A **subspace** of a vector space is a nonempty subset $W \subset V$ that is preserved by addition and scalar multiplication, i.e., $W + W \subset W$ and $k \cdot W \subset W$.*

In fact, equality $=$ holds instead of $\subset$, and the second condition implies $0 \in W$. Thus, W is also a vector space!

Example 2.19.

- $k^n = \{(a_1, \ldots, a_n) \mid a_i \in k\}$ *with component-wise addition and scalar multiplication.*
- $k^\infty = \{(a_i)_{i \in \mathbb{N}} \mid a_i \in k\}$ *(sequences in k) $\subset$ {sequences that are eventually zero}. (This corresponds to polynomials $k[x]$, and power series $k[\![x]\!]$).*
- *Given any set S, $k^S = \{$maps $f : S \to k\}$ ($k^\infty \iff$ case where $S = \mathbb{N}$).*
- $\{$*maps* $\mathbb{R} \to \mathbb{R}\} \supset \{$*continuous maps*$\} \supset \{$*differentiable maps* $\mathbb{R} \to \mathbb{R}\}$.

Let V be a vector space over k.

Definition 2.20. *Given $v_1, \ldots, v_n \in V$, the **span** of $v_1, \ldots, v_n$ is the smallest subspace of V that contains $v_1, \ldots, v_n$. Concretely,*

$$\mathit{span}(v_1, \ldots, v_n) = \{a_1 v_1 + \cdots + a_n v_n \mid a_i \in k\}.$$

In this case, we say that $v_1, \ldots, v_n$ span V.

Definition 2.21. *We say $v_1, \ldots, v_n$ are **linearly independent** if*

$$a_1 v_1 + \cdots + a_n v_n = 0 \implies a_1 = a_2 = \cdots = a_n = 0.$$

Equivalently, given $v_1, \ldots, v_n \in V$, we have a linear map $\phi : k^n \to V$, $(a_1, \ldots, a_n) \mapsto \sum a_i v_i$, such that $v_1, \ldots, v_n$ are linearly independent if and only if ϕ is injective, and $v_1, \ldots, v_n$ span V if and only if ϕ is surjective.

Definition 2.22. *The set $(v_1, \ldots, v_n)$ is a **basis** of V if they are linearly independent and span V.*

Then any element of V can be expressed uniquely as $\sum a_i v_i$ for some $a_i \in k$.

Example 2.23. *The vectors* $(0,1),(1,0)$ *form a basis of* k^2*. So do* $(1,1)$ *and* $(1,-1)$ *for most fields* k*. What if* $char(k) = 2$*?*

We will see soon that if V has a basis with n elements, then every basis of V has n elements. We say the **dimension** of V is $\dim(V) = n$.

One can also consider infinite-dimensional vector spaces.

Definition 2.24.

- *span*(S) *is the smallest subspace of* V *containing* S*, i.e.,*

$$\{a_1v_1 + \cdots + a_kv_k \mid k \in \mathbb{N}, a_i \in k, v_i \in S\}$$

  (all finite linear combinations of elements of S*).*

- *The elements of* S *are linearly independent if there are no finite linear relations:* $a_1v_1 + \cdots + a_kv_k = 0$ *(with* $a_i \in k, v_i \in S$*) implies* $a_1 = \cdots = a_k = 0$.

- S *is a basis of* V *if its elements are linearly independent and span* V.

Example 2.25. *The set* $\{1, x, x^2, x^3, \dots\}$ *is a basis of* $k[x]$.

Example 2.26. *Does* $k[\![x]\!]$ *have a basis? What is it?*

2.3 Linear Maps

Let V, W be vector spaces over k. A **homomorphism** of a vector space, or a **linear map**, $\varphi : V \to W$, is any map that is compatible with the operations:

$$\varphi(u+v) = \varphi(u) + \varphi(v), \quad \varphi(\lambda v) = \lambda\varphi(v) \quad \forall \lambda \in k, \forall u, v \in V.$$

Proposition 2.27. *The set of linear maps* $V \to W$ *is itself a vector space over* k*, denoted* $Hom(V, W)$.

Proof. Given $\varphi, \psi \in \mathrm{Hom}(V, W)$ for $\lambda \in k$, define:

1. $\varphi + \psi$ by $(\varphi + \psi)(v) = \varphi(v) + \psi(v)$,
2. $\lambda\varphi$ by $(\lambda\varphi)(v) = \lambda \cdot \varphi(v)$.

One can check that $\varphi + \psi$ and $\lambda\varphi$ are linear maps and these operations on $\mathrm{Hom}(V, W)$ satisfy the axioms of a vector space. □

Soon, we will see that if $\dim(V) = n$ and $\dim(W) = m$, then $\dim(\mathrm{Hom}(V, W)) = mn$. (In terms of bases for V and W, linear maps become $m \times n$ matrices!)

Now, let's consider the following question: How does the choice of the field k affect the discussion of vector spaces? Given a subfield $k' \subset k$ (e.g., $\mathbb{R} \subset \mathbb{C}$ or $\mathbb{Q} \subset \mathbb{R}$), a vector space over k can also be viewed as a vector space over k' by "restriction of scalars" (i.e., restricting scalar multiplication to the domain $k' \times V \subset k \times V$). In particular, k itself is a vector space over k'.

Example 2.28. *$\mathbb{C}$ is a vector space over itself (of dimension 1, with $\{1\}$ as a basis). It is also a vector space over $\mathbb{R}$ (of dimension 2, with basis $\{1, i\}$).*

If V, W are $\mathbb{C}$-vector spaces (and hence also $\mathbb{R}$-vector spaces), any $\mathbb{C}$-linear map is also $\mathbb{R}$-linear. However, the converse is not true:

$$\mathrm{Hom}_{\mathbb{C}}(V, W) \subsetneq \mathrm{Hom}_{\mathbb{R}}(V, W).$$

For example, complex conjugation $\mathbb{C} \to \mathbb{C}$, $z = a + bi \mapsto a - bi$, is $\mathbb{R}$-linear but not $\mathbb{C}$-linear. Thus, the choice of field k does indeed matter.

2.4 Basis and Dimension

Definition 2.29. *A vector space V is **finite-dimensional** if there exists a finite subset $\{v_1, \ldots, v_m\}$ that spans V, i.e., every element of V is a linear combination $\sum a_i v_i$.*

Lemma 2.30. *If $\{v_1, \ldots, v_m\}$ spans V, then a subset of $\{v_1, \ldots, v_m\}$ is a basis of V.*

Proof. If the set $\{v_1, \ldots, v_m\}$ is linearly independent, then it already forms a basis. Otherwise, there exists a non-trivial linear relation $\sum a_i v_i = 0$, where not all a_i are zero. We can solve for v_i as a linear combination of the others (if $a_i \neq 0$). Thus, we can remove v_i, and the remaining set $\{v_j \mid j \neq i\}$ still spans V. This process can be repeated until the remaining elements are linearly independent. □

Thus, every finite-dimensional vector space has a basis.

Lemma 2.31. *If $\{v_1, \ldots, v_m\}$ are linearly independent, then there exists a basis of V that contains $\{v_1, \ldots, v_m\}$.*

Proof. Let $\{w_1, \ldots, w_r\}$ be a spanning set for V. By induction, we can enlarge $\{v_1, \ldots, v_m\}$ to a basis of each subspace $W_j = \mathrm{span}(\{v_1, \ldots, v_m, w_1, \ldots, w_j\}) \subset V$ for $j = 0, \ldots, r$. For $j = 0$, $\{v_1, \ldots, v_m\}$ is already a basis of W_0. Assuming that $\{v_1, \ldots, v_m, w_{i_1}, \ldots, w_{i_k}\}$ forms a basis of $W_{j-1} = \mathrm{span}(\{v_1, \ldots, v_m, w_1, \ldots, w_{j-1}\})$, if $W_j \subset W_{j-1}$, then we already have a basis of W_j. Otherwise, $\{v_1, \ldots, v_m, w_{i_1}, \ldots, w_{i_k}, w_j\}$ is linearly independent and spans W_j. This process ultimately results in a basis of $W_r = V$ since $\{w_1, \ldots, w_r\}$ spans V. □

Proposition 2.32. *If $\{v_1, \ldots, v_m\}$ and $\{w_1, \ldots, w_n\}$ are bases of V, then $m = n$.*

Proof. We claim that there exists $j \in \{1, \ldots, n\}$ such that $\{v_1, \ldots, v_{m-1}, w_j\}$ is a basis. Indeed, the set $\{v_1, \ldots, v_{m-1}\}$ is linearly independent, but it does not span V, or else $v_m \in \mathrm{span}\{v_1, \ldots, v_{m-1}\}$ would give a linear relation $\sum_{i=1}^{m-1} a_i v_i - v_{m+1} = 0$. Therefore, there exists a j such that $w_j \notin \mathrm{span}\{v_1, \ldots, v_{m-1}\}$ (otherwise $w_1, \ldots, w_n$ could not span V).

Now, the set $\{v_1, \ldots, v_{m-1}, w_j\}$ is linearly independent. By expressing w_j in terms of the basis $\{v_1, \ldots, v_m\}$, we get $w_j = \sum_{i=1}^{m} a_i v_i$ (with $a_m \neq 0$), so that $v_m = \frac{1}{a_m}\left(w_j - \sum_{i=1}^{n-1} a_i v_i\right) \in \operatorname{span}(\{v_1, \ldots, v_{m-1}, w_j\})$, implying that $\{v_1, \ldots, v_{m-1}, w_j\}$ spans V and forms a basis.

Repeating this process allows us to exchange one v for one w at each step (without repeating the same w because the new w must be independent of the current basis). Ultimately, we will obtain a subset of $\{w_1, \ldots, w_n\}$ of size m that is also a basis. Since this subset is a basis, it must be the entire set $\{w_1, \ldots, w_n\}$, and thus $m = n$. □

Definition 2.33. *The **dimension** of V is the cardinality of any basis of V.*

Given a basis $(v_1, \ldots, v_n)$ of V, we can define a linear map $\varphi : k^n \to V$ by $(a_1, \ldots, a_n) \mapsto \sum a_i v_i$:

- Linear independence $\iff \varphi$ is injective, and spanning $V \iff \varphi$ is surjective, so φ is an isomorphism.
- Every finite-dimensional vector space over k is isomorphic to k^n, where $n = \dim(V)$ (and the basis gives a specific choice of such an isomorphism).

Given bases $(v_1, \ldots, v_n)$ of V and $(w_1, \ldots, w_m)$ of W, we can represent a linear map $\varphi \in \operatorname{Hom}(V, W)$ by an $m \times n$ matrix $A \in M_{m,n}$. This representation amounts to the following diagram:

$$\begin{array}{ccc} V & \xrightarrow{\varphi} & W \\ \uparrow & & \uparrow \\ \simeq\text{basis} & & \simeq\text{basis} \\ | & & | \\ k^n & \xrightarrow{A} & k^m \end{array}$$

We write $A = (a_{ij})_{1 \leq i \leq m, 1 \leq j \leq n} = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ a_{21} & & \\ \vdots & & \\ a_{m1} & \ldots & a_{mn} \end{pmatrix}$. The matrix A acts by multiplying column vectors:

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \mapsto A \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

Notation: $A = M(\varphi, (v), (w))$.

The entries of A are characterized by $\varphi(v_j) = \sum_{i=1}^{m} a_{ij} w_i$. In other words, the columns of A give the components of $\varphi(v_1), \ldots, \varphi(v_n)$ in the basis $\{w_1, \ldots, w_m\}$.

Representing any element $x \in V$ as

$$x = \sum_{i=1}^{n} x_i v_i \quad \Longleftrightarrow \quad \text{column vector} \quad X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},$$

and similarly for $y = \varphi(x) \in W$,

$$y = \sum y_i w_i \quad \Longleftrightarrow \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = AX.$$

As a memory aid, the isomorphism $k^n \xrightarrow{\sim} V$ given by the basis can be written symbolically as multiplication of row and column vectors:

$$(v_1, \ldots, v_n) \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \sum x_i v_i.$$

Thus, $\varphi((v_1, \ldots, v_n)X) = (w_1, \ldots, w_n)AX$, which corresponds to the commutative diagram.

This construction provides an isomorphism between the vector spaces $\mathrm{Hom}(V, W)$ and $M_{m,n}$. In particular, $\dim(\mathrm{Hom}(V, W)) = \dim(M_{m,n}) = mn$, and linear maps correspond to matrices.

What happens if we choose different bases for V and/or W? If we change the basis from $(v_1, \ldots, v_n)$ to $(v'_1, \ldots, v'_n)$, we can write $v'_j = \sum_{i=1}^{m} p_{ij} v_i$ and define an $n \times n$ matrix P whose j-th column gives the components of v'_j in the basis $(v_1, \ldots, v_n)$. Symbolically:

$$(v'_1, \ldots, v'_n) = (v_1, \ldots, v_n)P.$$

Thus,

$$(v'_1, \ldots, v'_n)X' = (v_1, \ldots, v_n)PX'.$$

This means that the element $x' \in V$, described by the column vector X' in the new basis, is described by $X = PX'$ in the old basis. More conceptually:

$$P = M(\mathrm{id}_v, (v'), (v)).$$

Similarly for W, we define the inverse transformation with:

$$Q = M(\mathrm{id}_w, (w), (w')).$$

Thus, we get:

$$\varphi((v'_1, \ldots, v'_n)X') = \varphi((v_1, \ldots, v_n)PX') = (w_1, \ldots, w_m)APX' = (w'_1, \ldots, w'_n)QAPX',$$

which gives $M(\varphi,(v'),(w')) = QAP$.

In particular, if $V = W$ and we change the basis for $\varphi \in \mathrm{Hom}(V,V)$, the matrices $A = M(\varphi,(v),(v))$ and $A' = M(\varphi,(v'),(v'))$ are related by $A' = P^{-1}AP$.

However, the point of linear algebra is to avoid dealing with these coordinate transformations and work with linear maps in a coordinate-free manner as much as possible.

2.5 Direct Sums and Products

Definition 2.34. *Given vector spaces V and W, the **direct sum** is defined as*

$$V \oplus W = V \times W = \{(v,w) \mid v \in V, w \in W\}.$$

For n vector spaces, the direct sum is given by

$$V_1 \oplus \cdots \oplus V_n = V_1 \times \cdots \times V_n = \{(v_1,\ldots,v_n) \mid v_i \in V_i\}.$$

For an infinite collection of vector spaces $(V_i)_{i\in I}$, the direct sum is

$$\bigoplus_{i\in I} V_i = \{(v_i)_{i\in I} \mid v_i \in V_i, \text{ and only finitely many } v_i \neq 0\},$$

which differs from the direct product:

$$\prod_{i\in I} V_i = \{(v_i)_{i\in I} \mid v_i \in V_i\}.$$

Example 2.35. *Consider the direct sum $\bigoplus_{n\in\mathbb{N}} k \simeq k[x]$ versus the direct product $\prod_{n\in\mathbb{N}} k \simeq k[\![x]\!]$.*

Definition 2.36. *Given subspaces $W_1,\ldots,W_n \subset V$ of some vector space V, we define:*

- *The **span** of $W_1,\ldots,W_n$ as $W_1+\cdots+W_n = \{w_1+\cdots+w_n \mid w_i \in W_i\} \subset V$.*
- *The subspaces $W_1,\ldots,W_n$ are **independent** if $w_1+\cdots+w_n = 0$ with $w_i \in W_i$ implies $w_i = 0$ for all i.*
- *If the subspaces $W_1,\ldots,W_n$ are independent and span V, we say that V has a **direct sum decomposition** of the form $V = W_1 \oplus \cdots \oplus W_n$.*

A relation to the previous notation: For each i, we have an inclusion map $W_i \hookrightarrow V$. We then assemble these into a linear map

$$\varphi : \bigoplus_{i=1}^{n} W_i \to V, \quad (w_1,\ldots,w_n) \mapsto \sum_{i=1}^{n} w_i.$$

The subspaces $W_1, \ldots, W_n$ span V if and only if φ is surjective, and they are independent if and only if φ is injective. If both conditions hold, then φ is an isomorphism

$$\bigoplus_{i=1}^{n} W_i \xrightarrow{\sim} V,$$

and we have a direct sum decomposition. In this case, we also have the dimension formula

$$\dim(V) = \sum_{i=1}^{n} \dim(W_i),$$

and a basis of V can be obtained by taking the union of the bases of $W_1, \ldots, W_n$.

For the case of two subspaces, we have the following:

- The subspaces W_1 and W_2 are independent if and only if $W_1 \cap W_2 = \{0\}$. Thus, $w_1 + w_2 = 0$ if and only if $w_1 = -w_2 \in W_1 \cap W_2$.
- The dimension of the sum is given by the formula

$$\dim(W_1 + W_2) = \dim(W_1) + \dim(W_2) - \dim(W_1 \cap W_2).$$

- The direct sum $V = W_1 \oplus W_2$ holds if and only if $W_1 \cap W_2 = \{0\}$ and $\dim(W_1) + \dim(W_2) = \dim(V)$.

Additionally, note that for any subspace $W \subset V$, there exists another subspace W' such that

$$W \oplus W' = V.$$

Note that W' is not necessarily unique. To find W', take a basis $\{w_1, \ldots, w_r\}$ of W, and complete it to a basis $\{w_1, \ldots, w_r, w'_1, \ldots, w'_s\}$ of V. Then, define $W' = \operatorname{span}(w'_1, \ldots, w'_s)$.

2.6 Rank and the Dimension Formula

Definition 2.37. *Given finite-dimensional vector spaces V and W, and a linear map $\varphi : V \to W$, we define the following:*

- *The **kernel** of φ is*

$$Ker(\varphi) = \{v \in V \mid \varphi(v) = 0\} \subset V.$$

- *The **image** of φ is*

$$Im(\varphi) = \{w \in W \mid \exists v \in V \text{ such that } \varphi(v) = w\} \subset W,$$

  which is a subspace of W.

- *The dimension of the image of φ, denoted by $dim(Im(\varphi))$, is called the **rank** of φ.*

Proposition 2.38 (The Dimension Formula)**.** *The following dimension formula holds:*

$$dim(Ker(\varphi)) + dim(Im(\varphi)) = dim(V).$$

Proof. Start by choosing a basis $\{u_1, \ldots, u_m\}$ for $\mathrm{Ker}(\varphi)$, and extend it to a basis $\{u_1, \ldots, u_m, v_1, \ldots, v_r\}$ of V. We claim that $\{\varphi(v_1), \ldots, \varphi(v_r)\}$ forms a basis for $\mathrm{Im}(\varphi)$. Indeed:

- If $w = \varphi(v) \in \mathrm{Im}(\varphi)$, then write $v = \sum a_i u_i + \sum b_j v_j$ and apply φ:

  $$\varphi(v) = \sum b_j \varphi(v_j),$$

  showing that $\{\varphi(v_j)\}$ spans $\mathrm{Im}(\varphi)$.

- If $\sum c_j \varphi(v_j) = 0$, then $\varphi\left(\sum c_j v_j\right) = 0$, which implies that $\sum c_j v_j \in \mathrm{Ker}(\varphi)$. Hence, we have

  $$\sum c_j v_j = \sum a_i u_i \quad \text{for some } a_i \in k.$$

  Since $\{u_1, \ldots, u_m, v_1, \ldots, v_r\}$ is linearly independent, this forces all $c_j = 0$ and $a_i = 0$. Therefore, the set $\{\varphi(v_j)\}$ is linearly independent.

Thus, $\{u_1, \ldots, u_m, v_1, \ldots, v_r\}$ is a basis of V, where $m = \dim(\mathrm{Ker}(\varphi))$ (since $\{u_1, \ldots, u_m\}$ is a basis of $\mathrm{Ker}(\varphi)$), and $r = \dim(\mathrm{Im}(\varphi)) = \mathrm{rank}(\varphi)$ (since $\{\varphi(v_1), \ldots, \varphi(v_r)\}$ forms a basis of $\mathrm{Im}(\varphi)$). Therefore, we conclude that

$$m + r = \dim(V).$$

□

Corollary 2.39. *Given a linear map $\varphi : V \to W$, there exist bases of V and W in which the matrix of φ has the form*

$$\left(\begin{array}{c|c} I & 0 \\ \hline 0 & 0 \end{array}\right),$$

where I is an $r \times r$ identity matrix, and $r = rank(\varphi)$.

Proof. Take a basis of V given by $\{v_1, \ldots, v_r, u_1, \ldots, u_m\}$, where $\{u_1, \ldots, u_m\}$ is a basis of $\mathrm{Ker}(\varphi)$. Extend the set $\{\varphi(v_1), \ldots, \varphi(v_r)\}$ (which is a basis of $\mathrm{Im}(\varphi)$) to a basis of W. □

Corollary 2.40. *For subspaces $W_1, W_2 \subset W$, the following holds:*

$$dim(W_1 + W_2) = dim(W_1) + dim(W_2) - dim(W_1 \cap W_2).$$

Proof. Consider the map from $V = W_1 \oplus W_2$ to W, defined by

$$\varphi(w_1, w_2) = w_1 + w_2.$$

Then, we have

$$\mathrm{Im}(\varphi) = W_1 + W_2,$$

and

$$\mathrm{Ker}(\varphi) = \{(u_1, \ldots, u_k) \mid u_i \in W_1 \cap W_2\} \simeq W_1 \cap W_2.$$

Thus, using the dimension formula, we get

$$\begin{aligned} \dim(\mathrm{Ker}(\varphi)) + \dim(\mathrm{Im}(\varphi)) &= \dim(W_1 \cap W_2) + \dim(W_1 + W_2) \\ &= \dim(V) \\ &= \dim(W_1) + \dim(W_2). \end{aligned}$$

□

2.7 Quotient and Dual Spaces

Definition 2.41. *Let V be a vector space over the field k, and let $U \subset V$ be a subspace. The* ***quotient space*** *$V/U = \{v + U\}$ is the space of cosets of U in V, with addition defined by*

$$(v + U) + (w + U) = (v + w) + U$$

and scalar multiplication given by

$$a(v + U) = av + U.$$

The linear map $V \xrightarrow{q} V/U$, defined by $v \mapsto v + U$, is surjective, with kernel equal to U. Hence, we obtain the exact sequence

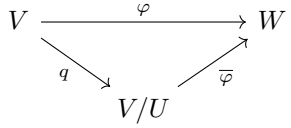
$$0 \to U \to V \to V/U \to 0.$$

By the dimension formula, we have

$$\dim(V/U) = \dim(V) - \dim(U).$$

Remark 2.42.

1. *Given a linear map $\varphi : V \to W$, if $U \subset Ker(q)$, then φ factors through V/U. Specifically, there exists a map $\overline{\varphi} : V/U \to W$ such that $\varphi = \overline{\varphi} \circ q$. This can be depicted in the following commutative diagram:*



Define $\overline{\varphi}(v+U) = \varphi(v)$, which is well-defined and independent of the choice of v in the coset.

Conversely, given $\overline{\varphi} \in Hom(V/U, W)$, the map $\varphi = \overline{\varphi} \circ q : V \to W$ satisfies $U \subset Ker(\varphi)$. Hence, we have the isomorphism

$$Hom(V,W)\big|_{U \subset Ker(\varphi)} \simeq Hom(V/U, W).$$

2. *There is a bijection between the set of subspaces of V containing U and the set of subspaces of V/U. Specifically, for $W \subset V$ with $W \supset U$, we have*

$$W \mapsto W/U = \{w + U \mid w \in W\}.$$

Conversely, for $\overline{W} \subset V/U$, we have

$$q^{-1}(\overline{W}) \subset V.$$

Thus, we see that $U = q^{-1}(0) \subset q^{-1}(\overline{W})$ since $0 \in \overline{W}$.

Definition 2.43. *The **dual vector space** V^* is the space of linear functionals on V, i.e., the set of linear maps $V \to k$. Formally, we define*

$$V^* = Hom(V, k) = \{\text{linear maps } \ell : V \to k\}.$$

Example 2.44. *If $V = k^n = \{(x_1, \ldots, x_n) \mid x_i \in k\}$, then any tuple $(a_1, \ldots, a_n)$ with $a_i \in k$ determines a map $\ell_a : k^n \to k$ defined by*

$$\ell_a(x_1, \ldots, x_n) = \sum_{i=1}^{n} a_i x_i.$$

Conversely, if e_i is the standard basis of k^n, then given a linear functional $\ell : k^n \to k$, define $a_i = \ell(e_i)$. Thus, for any vector $(x_1, \ldots, x_n)$, we have

$$\ell(x_1, \ldots, x_n) = \ell\left(\sum_{i=1}^{n} x_i e_i\right) = \sum_{i=1}^{n} a_i x_i.$$

Hence, $\ell = \ell_a$. Therefore, we conclude that

$$(k^n)^* = \{(a_1, \ldots, a_n) \mid a_i \in k\} \simeq k^n.$$

More generally, given a finite-dimensional vector space V with a basis $\{e_1, \ldots, e_n\}$, any linear map $\ell : V \to k$ is uniquely determined by its values on the basis vectors $\ell(e_i)$. Thus, we obtain an isomorphism

$$V^* \simeq k^n, \quad \ell \mapsto (\ell(e_1), \ldots, \ell(e_n)).$$

Equivalently, we can describe a basis for V^* consisting of the linear functionals $\{e_1^*, \ldots, e_n^*\}$, where $e_i^*(e_i) = 1$ and $e_i^*(e_j) = 0$ for $i \neq j$. Therefore, any linear functional ℓ can be written as

$$\ell = \sum_{i=1}^{n} \ell(e_i) e_i^*.$$

This is known as the **dual basis** of V^*.

However, there is no natural map from V to V^*, even though both spaces have similar bases. Each element of the dual basis e_i^* depends not just on e_i, but on all the basis vectors e_j. Thus, we cannot speak of "the dual of a vector."

On the other hand, there is a natural map called the **evaluation map**:

$$V \overset{ev}{\to} (V^*)^*, \quad v \mapsto ev_v : V^* \to k, \quad \ell \mapsto \ell(v).$$

If V is finite-dimensional, then by working in bases $\{e_1, \ldots, e_n\}$ for V, the dual basis $\{e_1^*, \ldots, e_n^*\}$ for V^*, and the double dual basis $\{e_1^{**}, \ldots, e_n^{**}\}$ for V^{**}, we see that

$$e_i^{**}(e_j^*) = e_j^*(e_i),$$

and thus $e_v(e_i) = e_i^{**}$. Hence, the evaluation map ev is an isomorphism.

Proposition 2.45. *If V is finite-dimensional, then $V \simeq V^{**}$, where the isomorphism is given by*

$$v \mapsto (\ell \mapsto \ell(v)).$$

When V is infinite-dimensional, the evaluation map $ev : V \to V^{**}$ is injective, but not an isomorphism. The reason is that if V has a basis $\{e_i\}_{i \in I}$, then every element of V can be uniquely written as $\sum_{i \in I} x_i e_i$ with only finitely many nonzero x_i. Hence, $V \simeq \bigoplus_{i \in I} k e_i$. For each choice of $(a_i)_{i \in I} \in \prod_{i \in I} k$, we can define a linear functional $\ell_a : V \to k$ by

$$\ell_a \left(\sum_{i \in I} x_i e_i \right) = \sum_{i \in I} x_i a_i,$$

which is a well-defined element of V^*. Thus, we have the isomorphism

$$V^* \simeq \prod_{i \in I} k,$$

which is a larger space than V. The linear functionals $\{e_i^*\}$, with $a_i = 1$ and $a_j = 0$ for all $j \neq i$, do not span V^*. One can complete this set to a basis using Zorn's Lemma. A similar enlargement occurs when passing from V^* to V^{**}.

2.8 Annihilators and Transposes

Definition 2.46. *The **annihilator** of a subspace $U \subset V$ is the set*

$$Ann(U) = \{\ell : V \to k \mid \ell_{|U} = 0\} \subset V^*.$$

Proposition 2.47. *$Ann(U)$ is a subspace of V^*.*

We now state the following properties:

- The map $V^* \to U^*$ is surjective with kernel equal to $\mathrm{Ann}(U)$, yielding the exact sequence

  $$0 \to \mathrm{Ann}(U) \to V^* \to U^* \to 0, \quad \ell \mapsto \ell_{|U}.$$

  This implies that $V^*/\mathrm{Ann}(U) \simeq U^*$.

- As seen above, the map

  $$\{\ell \in \mathrm{Hom}(V, k) \mid U \subset \mathrm{Ker}(\ell)\} \simeq \mathrm{Hom}(V/U, k)$$

  gives the isomorphism $\mathrm{Ann}(U) \simeq (V/U)^*$.

- Consequently, we have the dimension formula

  $$\dim(\mathrm{Ann}(U)) = \dim(V) - \dim(U).$$

  Check: The map φ^* is linear.

Definition 2.48. *Given a linear map $\varphi : V \to W$, the **transpose** of φ, denoted $\varphi^* : W^* \to V^*$, is defined as follows: for a linear functional $\ell : W \to k$, we define*

$$\varphi^*(\ell) = \ell \circ \varphi : V \to k.$$

Thus, $\varphi^* : W^* \to V^*$ is the map $\ell \mapsto \varphi^*(\ell) = \ell \circ \varphi$.

Check: Given a basis $\{e_i\}$ of V, the elements of V can be represented by column vectors X and row vectors Y. Applying a linear functional $\ell \in V^*$ to a vector $v \in V$ corresponds to the action $YX \in k$.

If $M(\varphi, (e_i), (f_j)) = A$, then $M(\varphi^*, (f_j^*), (e_i^*)) = A^T$, the transpose of the matrix A. To see why, consider the following: for any $\ell \in W^*$ and $v \in V$, we have the equation

$$\ell(\varphi(v)) = (\varphi^{**}(\ell))(v) = YAX,$$

where Y and X are row vectors corresponding to ℓ and v, respectively.

Now, if we view φ^* as an operation on row vectors, we can see that φ^* maps Y to YA. On the other hand, the dual basis provides a representation of elements of V^* and W^* as column vectors, which are the transposes of the corresponding row vectors. Thus, the claim follows: since $\varphi^*\ell$ as a column vector is $(YA)^T = A^T Y^T$, we conclude that $M(\varphi^*, (f_j^*), (e_i^*)) = A^T$.

Proposition 2.49. *In the finite-dimensional case:*

1. *φ is injective if and only if φ^* is surjective.*
2. *φ is surjective if and only if φ^* is injective.*

This follows from the following property:

Proposition 2.50. *In the finite-dimensional case, we have:*

1. *$Ker(\varphi^*) = Ann(Im(\varphi))$.*
2. *$Im(\varphi^*) = Ann(Ker(\varphi))$.*

Proof.

1. $\mathrm{Ker}(\varphi^*) = \mathrm{Ann}(\mathrm{Im}(\varphi))$: If $\ell \in \mathrm{Ann}(\mathrm{Im}(\varphi))$, then
$$\ell(\varphi(v)) = 0 \quad \forall v \in V,$$
which implies $\varphi^*(\ell) = \ell \circ \varphi = 0$. Hence, $\ell \in \mathrm{Ker}(\varphi^*)$.
2. $\mathrm{Im}(\varphi^*) = \mathrm{Ann}(\mathrm{Ker}(\varphi))$: If $\ell' = \varphi^*(\ell) \in \mathrm{Im}(\varphi^*)$, then $\ell' = \ell \circ \varphi$. Therefore, for any $v \in \mathrm{Ker}(\varphi)$, we have
$$\ell'_{|\mathrm{Ker}(\varphi)} = 0.$$
Thus, $\ell' \in \mathrm{Ann}(\mathrm{Ker}(\varphi))$, so $\mathrm{Im}(\varphi^*) \subset \mathrm{Ann}(\mathrm{Ker}(\varphi))$.

   By the dimension formula and the previous result, we have $\mathrm{rank}(\varphi^*) = \mathrm{rank}(\varphi)$. Therefore, the conclusion follows.

□

2.9 Linear Operators and Invariant Subspaces

Definition 2.51. *A **linear operator** on V (also called an **endomorphism** of V) is a linear map $\varphi : V \to V$.*

Notation: $\mathrm{End}(V) = \mathrm{Hom}(V, V)$.

When expressing $\varphi \in \mathrm{Hom}(V, V)$ using a basis, we use the same basis on both sides. If $A = M(\varphi, (e_i), (e_i))$ represents the matrix of φ in a given basis (e_i), then A transforms as $P^{-1}AP$ under a change of basis.

Note that if $\dim(V) \leq \infty$, the following are equivalent:

$$\begin{aligned} \varphi : V \to V \text{ is injective} &\iff \varphi \text{ is surjective} \\ &\iff \varphi \text{ is an isomorphism} \\ &\iff \mathrm{rank}(\varphi) = \dim(V). \end{aligned}$$

In this case, we say that φ is invertible, and its inverse $\varphi^{-1} : V \to V$ exists.

At this point, we can compose linear operators. If φ and ψ are operators, their composition is denoted by $\varphi \circ \psi$, and it is a map $v \to V$. We can also compose operators with themselves, writing $\varphi^n = \varphi \circ \varphi \circ \cdots \circ \varphi$ (with n factors), or apply polynomials to them. For a polynomial $p(x) = \sum a_n x^n$, we define $p(\varphi) = \sum a_n \varphi^n$, which is also a linear map $V \to V$. Additionally, $\mathrm{Hom}(V, V)$ is a (noncommutative) ring under composition.

Given vector spaces V_1, V_2 and linear operators $\varphi_i : V_i \to V_i$, we can define the operator $\varphi = \varphi_1 \oplus \varphi_2 : V_1 \oplus V_2 \to V_1 \oplus V_2$ on the direct sum $V = V_1 \oplus V_2$. The operator φ leaves the subspaces V_1 and V_2 invariant, meaning that $\varphi(V_i) \subset V_i$. In a basis of V such that $e_1, \ldots, e_m \in V_1$ and $e_{m+1}, \ldots, e_n \in V_2$, the matrix of φ is block diagonal:

$$\left(\begin{array}{c|c} \varphi_1 & 0 \\ \hline 0 & \varphi_2 \end{array}\right).$$

Conversely, if $V = V_1 \oplus V_2$ and $\varphi(V_i) \subset V_i$ for $i = 1, 2$, then the matrix of φ is of this block diagonal form. More generally, if $V_1 \subset V$ is invariant (i.e., $\varphi(V_1) \subset V_1$) but V_2 is not necessarily invariant, then the matrix of φ will be block triangular:

$$\left(\begin{array}{c|c} \varphi_1 & * \\ \hline 0 & * \end{array}\right).$$

Thus, a common strategy for studying $\varphi : V \to V$ is to search for invariant subspaces.

Proposition 2.52. *If $U \subset V$ is invariant and $\dim(U) = 1$ (so $U = k \cdot v$ for some $v \in V$), then necessarily $\varphi(v) = \lambda v$ for some $\lambda \in k$.*

2.10 Eigenvectors and Eigenvalues

Definition 2.53. *A **eigenvector** of a linear map $\varphi : V \to V$ is a nonzero vector $v \in V$ such that $\varphi(v) = \lambda v$ for some scalar $\lambda \in k$. The scalar λ is called the **eigenvalue** corresponding to v.*

If we can find a basis of V consisting entirely of eigenvectors of φ, then we say that φ is diagonalizable. In this case, we can express φ in a basis where its matrix representation is diagonal:

$$\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix},$$

with $\varphi(v_i) = \lambda_i v_i$ for each v_i.

This is the best outcome, but it is not always possible!

Example 2.54. *Consider the space $V = \mathbb{R}^2$. The matrix*

$$\begin{pmatrix} \lambda_1 & 0 \\ 0 & \mu \end{pmatrix}$$

has eigenvectors $(1, 0)$ and $(0, 1)$ (or any scalar multiples) with eigenvalues λ_1 and μ, respectively.

However, the matrix

$$\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

has only one eigenvector, $(1, 0)$ (up to scaling), with eigenvalue 1, since it is not diagonalizable.

Moreover, the matrix

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

has no eigenvectors.

Proposition 2.55. *Eigenvectors of a linear map $\varphi : V \to V$ corresponding to distinct eigenvalues are linearly independent.*

Proof. Let $v_1, v_2, \ldots, v_\ell$ be eigenvectors of φ with distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_\ell$, so that $\varphi(v_i) = \lambda_i v_i$ for each i. Assume there exists a nontrivial linear combination $a_1 v_1 + a_2 v_2 + \cdots + a_\ell v_\ell = 0$ where not all a_i are zero. Without loss of generality, suppose $a_1 \neq 0$.

Then, applying φ to the linear combination:

$$\varphi(a_1 v_1 + a_2 v_2 + \cdots + a_\ell v_\ell) = a_1 \lambda_1 v_1 + a_2 \lambda_2 v_2 + \cdots + a_\ell \lambda_\ell v_\ell = 0.$$

This is another linear relation. Subtracting λ_1 times the first term, we obtain a new relation:

$$a_1(\lambda_1 - \lambda_1)v_1 + \cdots + a_\ell(\lambda_\ell - \lambda_1)v_\ell = 0,$$

where all coefficients except a_1 remain nonzero since $\lambda_i \neq \lambda_1$ for $i \neq 1$. This contradicts the assumption that the linear combination had the fewest nonzero coefficients, completing the proof. □

Corollary 2.56. *The number of distinct eigenvalues of a linear map $\varphi \in Hom(V, V)$ is at most $n = dim(V)$. If equality holds, then φ is diagonalizable.*

Definition 2.57. *A field k is said to be **algebraically closed** if every non-constant polynomial $p(x) \in k[x]$ has at least one root in k, i.e., there exists an $\alpha \in k$ such that $p(\alpha) = 0$.*

If k is algebraically closed, by the division algorithm for polynomials, any polynomial $p(x)$ can be factored as:

$$p(x) = c(x - \alpha_1)(x - \alpha_2) \cdots (x - \alpha_d),$$

where $d = \deg(p)$ and $\alpha_i \in k$.

The fundamental theorem of algebra states that $\mathbb{C}$ is algebraically closed. However, the proof is not purely algebraic; we will discuss it in Part B of this course.

If k is not algebraically closed, there exists an algebraically closed extension field $\overline{k}$ containing k, constructed by adjoining roots of polynomials in $k[x]$.

Example 2.58. *For example, $\overline{\mathbb{R}} = \mathbb{C}$, while $\overline{\mathbb{Q}}$ is the field of all roots of polynomials with rational coefficients, i.e., $\overline{\mathbb{Q}} \subset \mathbb{C}$.*

Lemma 2.59. *If k is algebraically closed and V is a finite-dimensional vector space over k, then any linear map $\varphi : V \to V$ has at least one eigenvector, i.e., there exists $v \in V \setminus \{0\}$ and $\lambda \in k$ such that $\varphi(v) = \lambda v$.*

Proof. Let $n = \dim(V)$. Choose a nonzero vector $v \in V$. The set of vectors $v, \varphi(v), \varphi^2(v), \ldots, \varphi^n(v)$ must be linearly dependent, so there exist scalars $a_0, a_1, \ldots, a_n \in k$, not all zero, such that:

$$a_0 v + a_1 \varphi(v) + a_2 \varphi^2(v) + \cdots + a_n \varphi^n(v) = 0.$$

This is a polynomial equation in φ. Since k is algebraically closed, we can factor the polynomial:

$$a_0 + a_1 \varphi + a_2 \varphi^2 + \cdots + a_n \varphi^n = c(\varphi - \lambda_1)(\varphi - \lambda_2) \cdots (\varphi - \lambda_d),$$

where $c \neq 0$ and the $\lambda_i \in k$. Since the operator $c(\varphi - \lambda_1) \cdots (\varphi - \lambda_d)$ is not invertible, it must have a nontrivial kernel, which implies the existence of an eigenvector v with eigenvalue λ_i. □

Corollary 2.60. *Given a linear map $\varphi : V \to V$ over an algebraically closed field k, there exists a basis $(v_1, v_2, \ldots, v_n)$ of V such that the matrix of φ is upper triangular, i.e., for each k, the subspace $V_k = \mathrm{span}(v_1, \ldots, v_k)$ is invariant under φ.*

Proof. We prove this by induction on $\dim(V)$. If $\dim(V) = 1$, then any nonzero vector v_1 is mapped to a scalar multiple of itself, and any 1×1 matrix is trivially upper triangular.

Assume the result holds for $\dim(V) \leq n - 1$ and consider $\varphi : V \to V$ where $\dim(V) = n$. By the lemma, φ has at least one eigenvalue $\lambda \in k$. Let $U = \mathrm{Im}(\varphi - \lambda)$. Since $\varphi - \lambda$ has nontrivial kernel (eigenvectors for λ), $\dim(U) < \dim(V)$. Moreover, U is an invariant subspace under φ.

By the induction hypothesis, $\varphi|_U \in \mathrm{Hom}(U, U)$ is upper triangular, so there exists a basis $(u_1, \ldots, u_m)$ of U such that $\varphi|_U$ is upper triangular. Now, complete this basis to a basis $(u_1, \ldots, u_m, v_1, \ldots, v_k)$ of V. This gives the required upper triangular form for φ. □

Remark 2.61. *This proof also proceeds by induction, but with a different starting point. We begin with $V_0 = k \cdot v_0$, where v_0 is an eigenvector of φ, and define $U = V/V_0$ as the quotient space. Let $q : V \to U$ denote the quotient map. Since $\varphi(V_0) \subset V_0$, there exists a map $\overline{\varphi} : U \to U$ such that the following diagram commutes:*

$$\begin{array}{ccc} V & \xrightarrow{\varphi} & V \\ {\scriptstyle q}\downarrow & & \downarrow{\scriptstyle q}\,. \\ U & \xrightarrow[\overline{\varphi}]{} & U \end{array}$$

The commutativity of this diagram follows because $(q \circ \varphi)_{|V_0} = 0$, meaning that $q \circ \varphi = \overline{\varphi}$ on the quotient space U. Thus, the map φ factors through $V/V_0 = U$.

By the induction hypothesis, there exists a basis $\{u_1, \ldots, u_{n-1}\}$ of U such that $\overline{\varphi}(u_i) \in span(u_1, \ldots, u_i)$. Now, for each i, let $v_i \in V$ such that $q(v_i) = u_i$. Then we have:

$$q(\varphi(v_i)) \in span(u_1, \ldots, u_i),$$

which implies that $\varphi(v_i) \in span(v_0, v_1, \ldots, v_i)$, since $(v_0, \ldots, v_{n-1})$ is a basis of V.

Now, suppose we have $\varphi : V \to V$ and a basis $(v_1, ..., v_n)$ of V such that $M(\varphi) = A$ is upper-triangular, i.e. each $V_i = \text{span}(v_1, ..., v_i)$ is an invariant subspace of φ. Denote by $\lambda_i = a_{ii}$ the entries of E.

Lemma 2.62. *φ is invertible if and only if all the diagonal entries of A are nonzero.*

Proof. If all λ_i are nonzero, then φ is surjective, and hence an isomorphism. To see this, consider the following:

- Since $\varphi(v_1) = \lambda_1 v_1$ and $\lambda_1 \neq 0$, we have $v_1 \in \text{Im}(\varphi)$.
- Next, for $\varphi(v_2) = \lambda_2 v_2 + a_{12} v_1$ with $\lambda_2 \neq 0$, we can solve for v_2 as follows:

$$v_2 = \frac{1}{\lambda_2}(\varphi(v_2) - a_{12} v_1) \in \text{Im}(\varphi).$$

- Repeating this argument for each v_i, we conclude that $v_i \in \text{Im}(\varphi)$ for all i.

Thus, if all λ_i are nonzero, φ is surjective.

On the other hand, if any $\lambda_i = 0$, then $\varphi(V_i) \subseteq V_{i-1}$. In this case, the restriction $\varphi|_{V_i}$ has a nontrivial kernel, since:

$$\text{rank}(\varphi|_{V_i}) \leq \dim(V_{i-1}) < \dim(V_i).$$

Therefore, $\text{Ker}(\varphi|_{V_i}) \neq 0$, implying that φ is not invertible. □

Corollary 2.63. *The following are equivalent:*

1. *λ is an eigenvalue of φ.*
2. *$\varphi - \lambda$ is not invertible.*
3. *λ is a diagonal entry of the upper triangular matrix A representing φ.*

Proof. $1 \iff 2$: Eigenvectors correspond to the kernel of $\varphi - \lambda$.

$2 \iff 3$: By applying the lemma to $\varphi - \lambda$, we conclude that λ is a diagonal entry of the matrix A representing φ. □

2.11 Generalized Eigenvectors

We have just discussed linear operators $\varphi : V \to V$, their invariant subspaces ($U \subset V$ such that $\varphi(U) \subset U$), and eigenvectors ($v \neq 0$ such that $\varphi(v) = \lambda v$, i.e., $v \in \operatorname{Ker}(\varphi - \lambda I)$).

Over any field:

- Eigenvalues need not exist; eigenvectors corresponding to distinct λ are linearly independent.
- If there are $n = \dim V$ distinct eigenvalues, then φ is diagonalizable, meaning there exists a basis such that the matrix representation of φ is:
$$M(\varphi) = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

For algebraically closed fields, such as $\mathbb{C}$, we have:

- Every operator has at least one eigenvector.
- There exists a basis such that the matrix representation of φ is upper triangular:
$$M(\varphi) = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$
(This is equivalent to the fact that the subspaces $V_i = \operatorname{span}(v_1, \ldots, v_i)$ are all invariant).
- $\varphi - \lambda I$ is invertible $\iff \lambda \notin \{\lambda_1, \ldots, \lambda_n\}$, meaning that the diagonal entries of the matrix are the eigenvalues of φ.

Next, we explore the study of invariant subspaces and eigenvalues for linear operators over algebraically closed fields, particularly $\mathbb{C}$. This leads us to the Jordan normal form.

Recall that the $\operatorname{Ker}(\varphi) = \{v \in V \mid \varphi(v) = 0\}$.

Definition 2.64. *The **generalized kernel** of φ is given by:*

$$gker(\varphi) = \{v \in V \mid \exists m \geq 0 \text{ such that } \varphi^m(v) = 0\}.$$

The generalized kernel consists of all vectors that eventually get mapped to 0 by repeated applications of φ.

Proposition 2.65. *The sequence of kernels is nested:*

$$0 \subset Ker(\varphi) \subset Ker(\varphi^2) \subset \cdots$$

because $\varphi^m(v) = 0 \implies \varphi^{m+1}(v) = 0$. If $Ker(\varphi^m) = Ker(\varphi^{m+1})$, then the sequence becomes constant after that.

Proof. Since $\mathrm{Ker}(\varphi^m) = \varphi^{-1}(\mathrm{Ker}(\varphi^{m+1}))$, we have $\mathrm{Ker}(\varphi^m) = \mathrm{Ker}(\varphi^{m+1}) \implies \mathrm{Ker}(\varphi^{m+1}) = \mathrm{Ker}(\varphi^{m+2})$. □

Because the sequence of kernels stops increasing after at most $n = \dim V$ steps, we conclude that:

$$\mathrm{gker}(\varphi) = \mathrm{Ker}(\varphi^n).$$

Example 2.66. *Consider the linear operator $\varphi : k^2 \to k^2$ defined by $e_1 \mapsto 0$ and $e_2 \mapsto e_1$, represented by the matrix:*

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Then, $Ker(\varphi) = k \cdot e_1$, but $Ker(\varphi^2) = gker(\varphi) = k^2$.

Lemma 2.67. *If $gker(\varphi) = Ker(\varphi^m)$, then $V = Ker(\varphi^m) \oplus Im(\varphi^m)$.*

Proof. Let $v \in \mathrm{Im}(\varphi^m) \cap \mathrm{Ker}(\varphi^m)$. Then $v = \varphi^m(u)$ for some $u \in V$. Since $v \in \mathrm{Ker}(\varphi^m)$, we have:

$$\varphi^m(v) = \varphi^m(\varphi^m(u)) = \varphi^{2m}(u) = 0.$$

Thus, $u \in \mathrm{Ker}(\varphi^{2m})$, and since $\mathrm{Ker}(\varphi^{2m}) = \mathrm{Ker}(\varphi^m)$ (by assumption $\mathrm{gker}(\varphi) = \mathrm{Ker}(\varphi^m)$), we conclude that $u \in \mathrm{Ker}(\varphi^m)$.

Therefore, $v = \varphi^m(u) = 0$, implying that:

$$\mathrm{Im}(\varphi^m) \cap \mathrm{Ker}(\varphi^m) = \{0\}.$$

By the dimension formula, we conclude that:

$$V = \mathrm{Ker}(\varphi^m) \oplus \mathrm{Im}(\varphi^m).$$

□

Definition 2.68. *We say that φ is **nilpotent** if there exists some $m \geq 0$ such that $\varphi^m = 0$, i.e., $gker(\varphi) = V$.*

We can apply the same ideas to eigenspaces.

Definition 2.69. *A vector $v \in V$ is a **generalized eigenvector** of φ with generalized eigenvalue λ if $v \in gker(\varphi - \lambda I)$, i.e., there exists some $m \geq 0$ such that $(\varphi - \lambda I)^m v = 0$. The subspace $gker(\varphi - \lambda I)$ is called the **generalized eigenspace**.*

Definition 2.70. *The **multiplicity** of the eigenvalue λ is the dimension of the generalized eigenspace:*

$$V_\lambda = gker(\varphi - \lambda I) = Ker((\varphi - \lambda I)^n).$$

In a basis where the matrix of φ is triangular, this multiplicity is the number of times λ appears on the diagonal.

Proposition 2.71. *The generalized eigenspaces $V_\lambda = Ker(\varphi - \lambda I)^n$ and $W_\lambda = Ker(\varphi - \lambda I)^n$ are invariant subspaces of φ, and:*

$$V = V_\lambda \oplus W_\lambda.$$

Proof. Let $v \in V_\lambda$. Then, $(\varphi - \lambda I)^n v = 0$, so $\varphi(\varphi - \lambda I)^n v = 0$. Since $\varphi - \lambda I$ commutes with φ, this implies $(\varphi - \lambda I)^n \varphi v = 0$, meaning $\varphi(v) \in V_\lambda$. If $v = (\varphi - \lambda I)^n u \in W_\lambda$, then:

$$\varphi(v) = \varphi(\varphi - \lambda I)^n u = (\varphi - \lambda I)^n \varphi(u) \in \mathrm{Im}(\varphi - \lambda I)^n = W_\lambda.$$

The lemma above, applied to $\varphi - \lambda I$, implies that $V = \mathrm{Ker}((\varphi - \lambda I)^n) \oplus \mathrm{Im}((\varphi - \lambda I)^n)$. □

Proposition 2.72. *The subspaces $V_\lambda \subset V$ are independent: if $\sum v_i = 0$, with $v_i \in V_{\lambda_i}$ and λ_i distinct, then $v_i = 0$ for all i.*

Proof. Assume $\sum_{i=1}^{\ell} v_i = 0$, where $v_i \in V_{\lambda_i}$ and the λ_i's are distinct. We will show $v_i = 0$ for all i.

If $v_i \neq 0$, let $k \geq 0$ be the largest integer such that $(\varphi - \lambda_i I)^k v_i = w \neq 0$, where $(\varphi - \lambda_i I)^{k+1} v_i = 0$, implying $\varphi(w) = \lambda_i w$.

Next, observe that:

$$(\varphi - \lambda_\ell I)^n (\varphi - \lambda_{\ell-1} I)^n \cdots (\varphi - \lambda_1 I)^n (v_1 + \cdots + v_\ell) = 0$$

since $v_1 + \cdots + v_\ell = 0$. This expression simplifies to the sum of terms of the form:

$$(\varphi - \lambda_\ell I)^n \cdots (\varphi - \lambda_2 I)^n w = \prod_{j=2}^{\ell} (\lambda_1 - \lambda_j)^n w \neq 0,$$

and the remaining terms vanish:

$$(\varphi - \lambda_\ell I)^n \cdots (\varphi - \lambda_2 I)^n (\varphi - \lambda_1 I)^n (v_1 + \cdots + w_\ell) v_j = 0 \quad \text{for all} \quad j \geq 2.$$

This leads to a contradiction, hence $v_i = 0$ for all i. □

Proposition 2.73. *Let k be algebraically closed, and let V be a finite-dimensional vector space over k. If $\varphi : V \to V$, then V decomposes into the direct sum of the generalized eigenspaces of φ, i.e.,*

$$V = \bigoplus_\lambda V_\lambda.$$

Proof. We induct on $\dim(V)$. The result is clear for $\dim(V) = 1$. Assume the result holds for all vector spaces of dimension $n-1$, and consider the case where $\dim(V) = n$.

Since k is algebraically closed, φ has at least one eigenvalue λ_1. Let $V_{\lambda_1} = \operatorname{gker}(\varphi - \lambda_1 I) = \operatorname{Ker}((\varphi - \lambda_1 I)^n)$, and let $U = W_{\lambda_1} = \operatorname{Im}((\varphi - \lambda_1 I)^n)$. Both V_{λ_1} and U are invariant subspaces, and we have:

$$V = V_{\lambda_1} \oplus U.$$

Since $\dim(U) < \dim(V)$, by the induction hypothesis, U decomposes into generalized eigenspaces for $\varphi_{|U}$:

$$U = U_{\lambda_2} \oplus \cdots \oplus U_{\lambda_\ell},$$

where $\lambda_2, \ldots, \lambda_\ell$ are eigenvalues of $\varphi_{|U}$, which are eigenvalues of φ with an eigenvector in U.

Moreover, we have:

$$U_{\lambda_j} = \operatorname{Ker}((\varphi_{|U} - \lambda_j I)^n) = \operatorname{Ker}((\varphi - \lambda_j I)^n) \cap U = V_{\lambda_j} \cap U.$$

Note that $\varphi_{|U}$ does not have λ as an eigenvalue, since:

$$\operatorname{Ker}((\varphi - \lambda I)^n) \cap U = 0 \quad \Rightarrow \quad \lambda \notin \{\lambda_1, \ldots, \lambda_\ell\}.$$

Since the generalized eigenspaces V_{λ_j} contain the subspaces U_{λ_j} for all $j \geq 2$, we conclude that $V_{\lambda_1}, \ldots, V_{\lambda_\ell}$ span V, and they are independent. Hence, we have:

$$V = V_{\lambda_1} \oplus \cdots \oplus V_{\lambda_\ell}.$$

In fact, $V_{\lambda_j} = U_{\lambda_j}$ for all $j \geq 2$. In other words:

$$\operatorname{Im}((\varphi - \lambda_i I)^n) = \bigoplus_{j \neq i} \operatorname{Ker}((\varphi - \lambda_j I)^n).$$

□

The decomposition $V = \bigoplus V_{\lambda_i}$ gives us bases in which φ is represented by a block diagonal matrix:

$$\begin{pmatrix} \varphi_{\lambda_1} & & 0 \\ & \varphi_{\lambda_2} & \\ & & \ddots \\ 0 & & \varphi_{\lambda_\ell} \end{pmatrix}$$

Moreover, $\varphi_{|v_{\lambda_i}}$ can be represented by a triangular matrix in a suitable basis for v_{λ_i}. Since its only eigenvalue is λ_i, all the diagonal entries of this matrix are equal to λ_i. Thus,

$$\varphi \sim \begin{pmatrix} \lambda_1 & * & & & & & \\ & \ddots & & & & 0 & \\ 0 & \lambda_1 & & & & & \\ & & \lambda_2 & * & & & \\ & & & \ddots & & & \\ & & 0 & \lambda_2 & & & \\ & & & & \ddots & & \\ & & & & & \lambda_\ell & * \\ 0 & & & & & & \ddots \\ & & & & & 0 & \lambda_\ell \end{pmatrix}$$

We can further analyze the blocks

$$\begin{pmatrix} \lambda_i & & * \\ & \ddots & \\ 0 & & \lambda_i \end{pmatrix},$$

but this requires a more detailed study of nilpotent operators. Note that $\varphi_{|V_{\lambda_i}} - \lambda_i I$ is nilpotent!

2.12 Nilpotent Operators

Let $\varphi : V \to V$ be a nilpotent operator, i.e., $\varphi^m = 0$ for some $m \leq \dim V$. This result holds for any field. The goal is to find a "nice" basis for V with respect to φ.

Observe that if $\dim(V) = 2$, there are two cases: either $\varphi = 0$, or $\varphi^2 = 0$ but $\varphi \neq 0$. In the second case, let $\varphi \notin \ker(\varphi)$. Then, $\varphi(v) = u \in \ker(\varphi)$, so u and v are independent and form a basis. In this basis, the matrix of φ is

$$M(\varphi) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

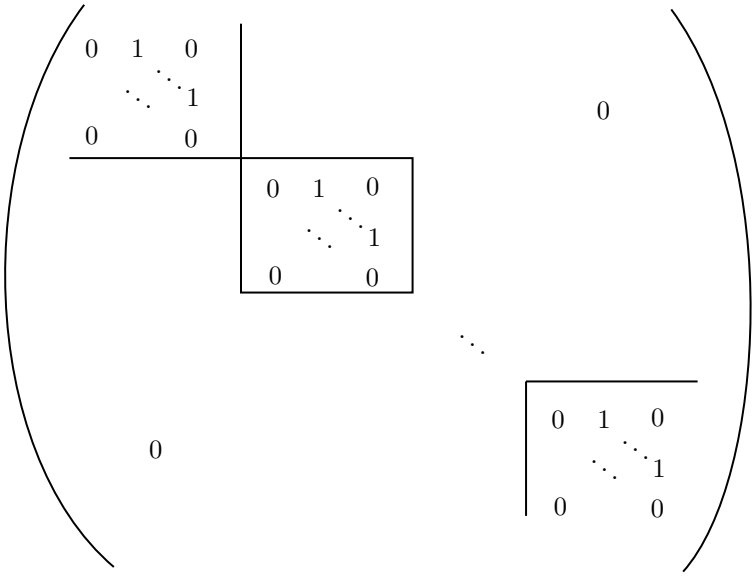
Jordan's method generalizes this to higher dimensions.

Proposition 2.74. *There exists a basis of V:*

$$\{\varphi^{m_1}(v_1), \varphi^{m_1-1}(v_1), \ldots, v_1, \ldots, \varphi^{m_k}(v_k), \ldots, v_k\}$$

where $\varphi^{m_i+1}(v_i) = 0$ for all i, in which the matrix of φ has the form:

$M(\varphi) \sim$



he block diagonals of $M(\varphi)$ form nilpotent Jordan blocks:

$$\begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{pmatrix},$$

where each basis vector maps to the previous one, and the first basis vector maps to 0.

Proof. Recall that the sequence of subspaces satisfies:

$$0 \subset \ker(\varphi) \subset \ker(\varphi^2) \subset \cdots \subset \ker(\varphi^m) = V.$$

We claim that if a subspace $U \subset \ker(\varphi^{k+1})$ satisfies $\ker(\varphi^k) \cap U = \{0\}$ for $k \geq 1$, then the restriction $\varphi_{|U}$ is injective, $\varphi(U) \subset \ker(\varphi^k)$, and $\ker(\varphi^{k+1}) \cap \varphi(U) = \{0\}$.

Indeed, for any $v \in U$, if $v \neq 0$, we have $\varphi^k(v) \neq 0$ and $\varphi^{k+1}(v) = 0$. In particular, $\varphi(v) \neq 0$, meaning $\ker(\varphi_{|U}) = \{0\}$, so $\varphi_{|U}$ is injective. Also, since $\varphi^k(\varphi(v)) = 0$, we have $\varphi(v) \in \ker(\varphi^k)$. Furthermore, $\varphi^{k-1}(\varphi(v)) = 0$ implies $\varphi(v) \in \ker(\varphi^{k-1})$, and $\varphi^{k-1}(\varphi(v)) = \varphi^k(v) \neq 0$, so $\varphi(v) \notin \ker(\varphi^{k-1})$.

Now, let U_m be a subspace such that $\ker(\varphi^m) = V = \ker(\varphi^{m-1}) \oplus U_m$. Choose a basis $(v_{m,1}, \dots, v_{m,k_m})$ for U_m, which will yield Jordan blocks of size m. We can extend this to a basis of V by adding vectors $v_{m,1}, \dots, v_{m,k_m}$ and letting U_m be their span.

By the claim, the vectors $v_{m-1,1} = \varphi(v_{m,1}, \dots, v_{m-1,k_m}) = \varphi(v_{m,k_m})$ are linearly independent, and their span is contained in $\ker(\varphi^{m-1})$ but independent of $\ker(\varphi^{m-2})$.

Starting from a basis of $\ker(\varphi^{m-2})$, add vectors $v_{m,1}, \dots, v_{m-1,k_m}$ and complete to a basis of $\ker(\varphi^{m-1})$ by adding other vectors $v_{m-1,k_m+1}, \dots, v_{m-1,k_{m-1}}$, yielding blocks of size $m-1$. Let $U_{m-1} = \operatorname{span}(v_{m-1,1}, \dots, v_{m-1,k_{m-1}})$. Then, $\ker(\varphi^{m-1}) = \ker(\varphi^{m-2}) \oplus U_{m-1}$.

We continue this process for all j, eventually constructing a basis of $V = U_1 \oplus \cdots \oplus U_m$. By rearranging the vectors as $(v_{1,1}, \dots, v_{m,1}, v_{1,2}, \dots)$, we obtain the desired basis. □

We can now combine our results to arrive at the following theorem:

Proposition 2.75. *Let V be a finite-dimensional vector space over an algebraically closed field k, and let $\varphi \in Hom(V, V)$. Then, there exists a basis of V in which the matrix of φ is block-diagonal, with each block being a Jordan block.*

Remark 2.76.

1. *φ is diagonalizable if and only if all the blocks have size 1.*
2. *The eigenvalues of φ are precisely the values λ that appear in the Jordan blocks. There may be several blocks with the same λ; their direct sum is the generalized eigenspace V_λ.*
3. *Here is an outline of the proof: We have seen that $V = \bigoplus V_\lambda$, where the V_λ are the generalized eigenspaces. Now, for each λ, the restriction $\varphi_{|V_\lambda} - \lambda I$ is nilpotent, and so V_λ can be decomposed into nilpotent Jordan blocks.*

2.13 Characteristic Polynomial

Let k be an algebraically closed field, and let $\varphi : V \to V$ be a linear map, where $V \bigoplus_{i=1}^{\ell} V_{\lambda_i}$ is a generalized eigenspace decomposition of V. For each i, define $n_i = \dim(V_{\lambda_i})$ to be the **multiplicity** of λ_i, with the condition that

$\sum n_i = \dim(V)$, and let m_i denote the nilpotency order of $(\varphi_{|V_{\lambda_i}} - \lambda_i \mathrm{Id})$, i.e., the smallest integer m_i such that $V_{\lambda_i} = \mathrm{Ker}((\varphi - \lambda_i I)^{m_i})$.

From the above, it follows that $m_i \leq n_i$, and V_{λ_i} is diagonalizable if and only if $m_i = 1$ for all i.

Definition 2.77. *The **characteristic polynomial** of φ is given by*

$$\chi_\varphi(x) = \prod_{i=1}^{\ell} (x - \lambda_i)^{n_i}.$$

The usual definition of the characteristic polynomial, once the determinant is defined, is

$$\chi_\varphi(x) = \det(xI - \varphi).$$

Manifestly, in a basis where $M(\varphi)$ is triangular (or Jordan normal form), we have

$$M(xI - \varphi) = \begin{pmatrix} x - \lambda_1 & & * \\ & \ddots & \\ & & x - \lambda_n \end{pmatrix}.$$

This is equivalent to the determinant definition, although any basis can be used to compute the determinant.

The significance of the characteristic polynomial is that, given the matrix representation A of φ in any basis, we can compute $\chi(x) = \det(xI - A) \in k[x]$, which allows us to find the roots (i.e., the eigenvalues) and their multiplicities (i.e., the dimensions of the generalized eigenspaces). This result holds even for non-algebraically closed fields k, although there is no guarantee that $\chi(x)$ has any roots in k.

Definition 2.78. *The **minimal polynomial** of φ is given by*

$$M_\varphi(x) = \prod_{i=1}^{\ell} (x - \lambda_i)^{m_i}.$$

The significance of the minimal polynomial is as follows: $(\varphi - \lambda_i)^k = 0$ on the generalized eigenspace V_{λ_i} if and only if $k \geq m_i$, and it is invertible on the other generalized eigenspaces. Hence, $M_\varphi(\varphi)$ is the simplest polynomial in φ that annihilates all V_{λ_i}'s, and therefore annihilates the entire space $V = \bigoplus_{i=1}^{\ell} V_{\lambda_i}$.

Thus, we have $M_\varphi(\varphi) = 0$, and for any polynomial $p \in k[x]$, $p(\varphi) = 0 \in \mathrm{Hom}(U, V)$ if and only if M_φ divides p. Since the nilpotency order m_i is always less than or equal to $\dim(V_{\lambda_i}) = n_i$, it follows that M_φ divides χ_φ. Therefore, we obtain the following result:

Proposition 2.79 (Cayley-Hamilton Theorem)**.**

$$\chi_\varphi(\varphi) = 0.$$

This result also holds for non-algebraically closed fields k, by passing to an algebraic closure. An example is given below.

3 Linear Algebra II

3.1 Real Operators

We now discuss operators on finite-dimensional real vector spaces. Let V be a real vector space and $\varphi : V \to V$ a linear operator. Since $\mathbb{R}$ is not algebraically closed, φ might not have eigenvalues, and it is not always possible to put φ into triangular or Jordan form. Nevertheless, every real operator has an invariant subspace of dimension 1 or 2. A common approach to address this issue is to work over $\mathbb{C}$, which is algebraically closed.

Definition 3.1. *The **complexification** of V is the vector space $V_{\mathbb{C}} = V \times V = \{v + iw \mid v, w \in V\}$, with addition defined by*

$$(v_1 + iw_1) + (v_2 + iw_2) = (v_1 + v_2) + i(w_1 + w_2),$$

and scalar multiplication defined by

$$(a + ib)(v + iw) = (av - bw) + i(bv + aw),$$

for all $a, b \in \mathbb{R}$.

This is a $\mathbb{C}$-vector space of dimension n. If $(e_1, \ldots, e_n)$ is a basis of V over $\mathbb{R}$, then $(e_1, \ldots, e_n)$ is also a basis of $V_{\mathbb{C}}$ over $\mathbb{C}$.

Given $\varphi : V \to V$ as an $\mathbb{R}$-linear operator, we can extend it to a $\mathbb{C}$-linear operator $\varphi_{\mathbb{C}} : V_{\mathbb{C}} \to V_{\mathbb{C}}$ by defining

$$\varphi_{\mathbb{C}}(v + iw) = \varphi(v) + i\varphi(w).$$

If $(e_1, \ldots, e_n)$ is a basis of V, the matrix representation of $\varphi_{\mathbb{C}}$ is the same as that of φ. However, now $\varphi_{\mathbb{C}}$ is guaranteed to have an eigenvector, along with generalized eigenspaces, Jordan form, and more.

Let $v' = v + iw$ be an eigenvector of $\varphi_{\mathbb{C}}$ corresponding to the eigenvalue $\lambda \in \mathbb{C}$, so that $\varphi_{\mathbb{C}}(v') = \lambda v'$. There are two possible cases:

- If $\lambda \in \mathbb{R}$, then

  $$\varphi_{\mathbb{C}}(v+iw) = \varphi(v)+i\varphi(w) = \lambda v+i\lambda w \implies v = \mathrm{Re}(v') \quad \text{and} \quad w = \mathrm{Im}(v')$$

  are eigenvectors of φ corresponding to the same eigenvalue λ (provided that either v or w is nonzero). The multiplicity of λ for φ does not necessarily need to be even.

- If $\lambda = a + ib \notin \mathbb{R}$, then

  $$\varphi_{\mathbb{C}}(v + iw) = (a + ib)(v + iw) \implies \varphi_{\mathbb{C}}(v - iw) = (a - ib)(v - iw),$$

  which follows by comparing real and imaginary parts. Hence, the complex conjugate $\overline{v'} = v - iw$ is also an eigenvector of $\varphi_{\mathbb{C}}$ with eigenvalue $\overline{\lambda}$. Consequently, v and w are linearly independent, and they span a 2-dimensional

invariant subspace $U \subset V$ where

$$\varphi(v) = av - bw, \quad \varphi(w) = bv + aw.$$

The matrix representation of $\varphi_{|U}$ in the basis (v, w) is

$$M(\varphi_{|U}, (v, w)) = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}.$$

Further analysis, such as the study of block triangular decompositions of φ, can be pursued starting from $\varphi_{\mathbb{C}}$.

3.2 Interlude: Category Theory

Definition 3.2. *A **category** is a collection of objects, and for each pair of objects, a collection of **morphisms** $Mor(A, B)$, along with an operation called **composition** of morphisms:*

$$Mor(A, B) \times Mor(B, C) \to Mor(A, C), \quad f, g \mapsto g \circ f,$$

such that the following hold:

1. *Every object A has an **identity morphism** $id_A \in Mor(A, A)$ such that for all $f \in Mor(A, B)$,*

$$f \circ id_A = id_B \circ f = f.$$

2. *Composition is **associative**:*

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

Example 3.3. *Examples of categories include:*

1. *The category of sets, Sets, where $Mor(A, B)$ is the set of all maps $A \to B$.*
2. *The category $Vect_k$, consisting of finite-dimensional vector spaces over k, with morphisms being linear maps.*
3. *The category of groups, Grp, where the morphisms are group homomorphisms.*
4. *The category of topological spaces, Top, with continuous maps as morphisms.*

Definition 3.4. *A morphism $f \in Mor(A, B)$ is an **isomorphism** if there exists a morphism $g \in Mor(A, B)$ (called the inverse isomorphism) such that*

$$g \circ f = id_A \quad \text{and} \quad f \circ g = id_B.$$

The following properties hold:

- The inverse of f, if it exists, is unique.
- id_A is an isomorphism.
- If f is an isomorphism, then f^{-1} is also an isomorphism.
- If f and g are isomorphisms, then $g \circ f$ is an isomorphism.

Thus, the automorphisms of an object A, denoted $\mathrm{Aut}(A) = \{\text{isomorphisms } A \to A\} \subset \mathrm{Mor}(A, A)$, form a group.

Isomorphic objects have isomorphic automorphism groups. Specifically, an isomorphism $f \in \mathrm{Mor}(A, B)$ induces an isomorphism of groups $c_f : \mathrm{Aut}(A) \to \mathrm{Aut}(B)$, defined by

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

Example 3.5.

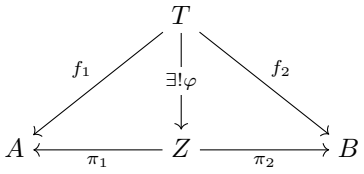
1. *In Sets, if A is a finite set with n elements, then $Aut(A) = \{$bijections $A \to A\} \cong S_n$, the symmetric group on n elements.*
2. *If V is an n-dimensional vector space over k, then $Aut(V) \cong GL_n(k)$, the group of invertible $n \times n$ matrices.*

Now, we discuss products and sums in categories:

Definition 3.6. *Given objects A and B in a category $\mathcal{C}$, a **product** $A \times B$ is an object Z of $\mathcal{C}$ along with two maps $\pi_1 : Z \to A$ and $\pi_2 : Z \to B$ such that for all objects $T \in Ob\mathcal{C}$, and for all morphisms $f_1 \in Mor(T, A)$ and $f_2 \in Mor(T, B)$, there exists a unique morphism $\varphi \in Mor(T, Z)$ such that*

$$\pi_1 \circ \varphi = f_1 \quad \text{and} \quad \pi_2 \circ \varphi = f_2.$$

This is represented as:



Example 3.7.

- *In Sets, the product $A \times B$ is the usual Cartesian product, with π_1 and π_2 being the projection maps. Given $f_1 : T \to A$ and $f_2 : T \to B$, we define*

$$T \to A \times B, \quad t \mapsto (f_1(t), f_2(t)).$$

- *In $Vect_k$, the product is $A \oplus B$ (so a sum is treated as a product in this context), with i_1 and i_2 being the inclusion maps $A \to A \oplus B$ and $B \to A \oplus B$, respectively. We define the morphism*

$$\varphi : T \to A \oplus B, \quad (a, b) \mapsto f_1(a) + f_2(b).$$

Next, we discuss functors:

Definition 3.8. *Let $\mathcal{C}$ and $\mathcal{D}$ be categories. A (covariant) **functor** $F : \mathcal{C} \to \mathcal{D}$ is an assignment that:*

- *assigns to each object X in $\mathcal{C}$, an object $F(X)$ in $\mathcal{D}$,*
- *assigns to each morphism $f \in Mor_{\mathcal{C}}(X, Y)$, a morphism $F(f) \in Mor_{\mathcal{D}}(F(X), F(Y))$,*

such that:

1. $F(id_X) = id_{F(X)}$,
2. $F(g \circ f) = F(g) \circ F(f)$.

Example 3.9. *Examples of functors include:*

1. *The **forgetful functor**, which takes a group, topological space, etc., to its underlying set.*
2. *For vector spaces, given a vector space V, the functor $F : W \mapsto Hom(V, W)$ assigns to each vector space W, the set of linear maps $Hom(V, W)$. If $f : W \to W'$ is linear, the induced map is*

$$Hom(V, W) \overset{F(f)}{\to} Hom(V, W'), \quad a \mapsto f \circ a.$$

   This gives a functor $Vect_k \to Vect_k$, denoted $Hom(V, \cdot)$.

3. *The **complexification functor** $Vect_{\mathbb{R}} \to Vect_{\mathbb{C}}$, which sends each vector space V over $\mathbb{R}$ to its complexification $V_{\mathbb{C}}$, and each morphism φ to $\varphi_{\mathbb{C}}$.*
4. *The functor*

$$Sets \to Groups, \quad X \mapsto \langle X \rangle,$$

   which sends a set X to the free group generated by X. For example, $F(\{a, b\}) = \langle a, b \rangle$, the free group on two generators.

A contravariant functor is a functor where the direction of morphisms is reversed.

Definition 3.10. *A **contravariant functor** is a functor such that for $f \in Mor_C(X, Y)$, we map f to $F(f) \in Mor_D(F(Y), F(X))$, and $F(g \circ f) = F(f) \circ F(g)$.*

Example 3.11. *On $Vect_k$, the dual functor $V \mapsto V^*$ is contravariant.*

Finally, we consider natural transformations:

Definition 3.12. *Given two functors $F, G : \mathcal{C} \to \mathcal{D}$, a **natural transformation** t from F to G consists of, for each object $X \in Ob\mathcal{C}$, a morphism $t_X \in Mor_{\mathcal{D}}(F(X), G(X))$ such that for all objects $X, Y \in Ob\mathcal{C}$, and for all*

morphisms $f \in Mor_{\mathcal{C}}(X, Y)$, the following diagram commutes:

$$\begin{array}{ccc} F(X) & \xrightarrow{t_X} & G(X) \\ {\scriptstyle F(f)}\downarrow & & \downarrow{\scriptstyle G(f)} \\ F(Y) & \xrightarrow{t_Y} & G(Y) \end{array}$$

Example 3.13. *The functor $Vect_k \to Sets$ that sends a vector space to its underlying set has natural transformations, which allow for compatibility between the functors.*

3.3 Bilinear Forms

Definition 3.14. *A **bilinear form** on a vector space V over a field K is a map $b : V \times V \to K$ that is linear in each variable: for all $u, v, w \in V$ and $\lambda \in K$, we have*

$$b(\lambda v, w) = b(v, \lambda w) = \lambda b(v, w), \quad b(u+v, w) = b(u, w)+b(v, w), \quad b(u, w+v) = b(u, w)+b(u, v).$$

Note that this is not a linear map $V \times V \to K$. Indeed, for a bilinear form, we have

$$b(\lambda(v, w)) = b(\lambda v, \lambda w) = \lambda^2 b(v, w) \neq \lambda b(v, w).$$

Definition 3.15. *We say b is **symmetric** if $b(v, w) = b(w, v)$ for all $v, w \in V$, and **skew-symmetric** if $b(v, w) = -b(w, v)$.*

Example 3.16.

- *The usual dot product on K^n, $(v, w) \mapsto \sum_{i=1}^{n} v_i w_i$, is symmetric.*
- *The map $b : K^2 \times K^2 \to K$, $b((x_1, x_2), (y_1, y_2)) = x_1y_2 - x_2y_1 = \det \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \end{pmatrix}$, is skew-symmetric.*

Given a bilinear map $b : V \times V \to K$, we can define a linear map $\varphi_b : V \to V^*$ by $\varphi_b(v) = b(v, \cdot) \in V^*$, which maps $w \in V$ to $b(v, w) \in K$. Conversely, a linear map $\varphi : V \to V^*$ determines a bilinear form $b(v, w) = (\varphi(v))(w)$. This defines a bijection

$$B(V) \xrightarrow{\sim} \mathrm{Hom}(V, V^*).$$

Definition 3.17. *The **rank** of the bilinear form $b : V \times V \to K$ is the rank of $\varphi_b : V \to V^*$, i.e., $dim(Im(\varphi_b))$. If φ_b is an isomorphism, we say that b is **nondegenerate**.*

For a given vector space V, the space of bilinear forms $B(V) = \{$bilinear forms $V \times V \to K\}$ is itself a vector space over K. What is its dimension?

If we choose a basis $\{e_1, \ldots, e_n\}$ for V, it is enough to specify $b(e_i, e_j)$ for all i, j in order to determine b. By bilinearity, we have

$$b\left(\sum_i x_i e_i, \sum_j y_j e_j\right) = \sum_{i,j} x_i y_j b(e_i, e_j).$$

Thus, the values of $b(e_i, e_j)$ can be chosen freely. For example, a basis of $B(V)$ is given by the set $(b_{k\ell})_{1\leq k\leq n, 1\leq \ell\leq n}$ such that

$$b_{k\ell}(e_i, e_j) = \begin{cases} 1 & \text{if } (i,j) = (k,\ell), \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, we have $\dim(B(V)) = (\dim(V))^2$, consistent with the isomorphism $B(V) \simeq \mathrm{Hom}(V, V^*)$. The map $b \mapsto \varphi_b$ is an isomorphism of vector spaces.

Given a basis $\{e_1, \ldots, e_n\}$ of V, the bilinear form $b : V \times V \to K$ is represented by an $n \times n$ matrix $A = (a_{ij})$ where $a_{ij} = b(e_i, e_j)$. We have

$$b\left(\sum_i x_i e_i, \sum_j y_j e_j\right) = \sum_{i,j} x_i y_j b(e_i, e_j) = (x_1, \ldots, x_n) A \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

so in terms of column vectors, we write $b(X, Y) = X^T A Y$.

Remark 3.18. *The isomorphism*

$$B(V) \xrightarrow{\sim} Hom(V, V^*) \quad \textit{given by} \quad b \mapsto \varphi_b$$

is natural in the sense that we have contravariant functors $V \mapsto B(V)$ and $V \mapsto Hom(V, V^*)$. For a morphism $f : V \to W$, we have

$$B(f) : B(W) \to B(V), \quad b(\cdot, \cdot) \mapsto b(f(\cdot), f(\cdot)),$$

and

$$Hom(W, W^*) \to Hom(V, V^*), \quad \varphi \mapsto f^* \circ \varphi \circ f.$$

The isomorphisms $B(V) \xrightarrow{\sim} Hom(V, V^*)$ define a natural transformation between these functors.

Definition 3.19. *If $S \subset V$ is a subspace of a vector space equipped with a bilinear form $b : V \times V \to K$, we define its* ***orthogonal complement*** *as the vector space*

$$S^\perp = Ann(\varphi_b(S)) = \{b(v, \cdot) \mid v \in S\} \subset V^*.$$

Equivalently, $S^\perp = \mathrm{Ann}(\varphi_b(S))$, where $\mathrm{Ann}(\varphi_b(S)) \subset V$ consists of the vectors on which all the linear forms in $\varphi_b(S)$ vanish. This is most useful when B is symmetric or skew-symmetric. Otherwise, we must be cautious about left/right orthogonality.

Proposition 3.20. *If b is nondegenerate, then $dim(S^\perp) = dim(V) - dim(S)$. Otherwise, $dim(S^\perp) = dim(V) - dim(\varphi_b(S))$.*

Example 3.21.

- *For $V = \mathbb{R}^n$ with the standard dot product, $V = S \oplus S^\perp$ is the "usual" orthogonal complement, since $S \cap S^\perp = \{0\}$ (if $v \in S \cap S^\perp$, then $b(v,v) = 0 \implies v = 0$) and $dim(S) + dim(S^\perp) = dim(V)$.*
- *For $b : K^2 \times K^2 \to K$, $b((x_1,x_2),(y_1,y_2)) = x_1y_2 - x_2y_1$, if $S \subset K^2$ is a 1-dimensional subspace spanned by any nonzero vector v, we have $S^\perp = S$ (because $b(v,w) = 0 \iff \det(v,w) = 0 \iff w \in K \cdot v = S$).*

3.4 Inner Product Spaces

Definition 3.22. *An **inner product space** is a vector space V over $\mathbb{R}$ together with a symmetric positive definite bilinear form $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$.*

Remark 3.23.

- ***Symmetric**: $\langle u, v \rangle = \langle v, u \rangle$*
- ***Positive definite**: $\langle u, u \rangle \geq 0$ for all $u \in V$, and $\langle u, u \rangle = 0$ if and only if $u = 0$.*

This definition is meaningful only over an ordered field, so the condition $\langle v, v \rangle \geq 0$ makes sense. In practice, this means the field $\mathbb{R}$. We cannot define an inner product over $\mathbb{C}$ in the same way, because $\langle iv, iv \rangle = i^2 \langle v, v \rangle = -\langle v, v \rangle$, which breaks the positivity condition for a bilinear form. However, there is a workaround: observe that $|\lambda|^2 \geq 0$ for all $\lambda \in \mathbb{C}$, which allows us to define Hermitian forms.

Definition 3.24. *Let V be a vector space over $\mathbb{C}$. A **Hermitian form** is a map $h : V \times V \to \mathbb{C}$ which is linear in the second variable and conjugate linear (or "complex antilinear") in the first variable:*

$$h(\lambda v, w) = \overline{\lambda} h(v,w) \forall \lambda \in \mathbb{C} \text{ vs. } h(v, \lambda w) = \lambda h(v,w)$$
$$h(v_1 + v_2, w) = h(v_1, w) + h(v_2, w) \text{ vs. } h(v, w_1 + w_2) = h(v, w_1) + h(v, w_2)$$

and conjugate symmetric: $h(v,w) = \overline{h(w,v)}$.

We then study $\mathbb{C}$-vector spaces with a Hermitian inner product, which is a positive-definite Hermitian form.

Let $\varphi : V \to V^*, v \mapsto \langle v, \cdot \rangle$ be the linear map corresponding to the inner product $\langle \cdot, \cdot \rangle$. If $\langle \cdot, \cdot \rangle$ is positive definite, then φ is injective (since for all $v \neq 0$, $\varphi(v) \neq 0$, implying $v\varphi(v) > 0$). Thus, assuming $\dim(V) < \infty$, φ is an isomorphism $V \xrightarrow{\sim} V^*$, i.e., $\langle \cdot, \cdot \rangle$ is nondegenerate. Note that the converse of this is false.

Proposition 3.25. *If V is a finite-dimensional inner product space and $S \subset V$ is a subspace, then $V = S \oplus S^\perp$.*

Proof. Since $\langle \cdot, \cdot \rangle$ is nondegenerate, we have $\dim(S^\perp) = \dim(V) - \dim(S)$. Furthermore, since $\langle \cdot, \cdot \rangle$ is positive definite, $v \in S \cap S^\perp$ implies $\langle v, v \rangle = 0$, and hence $v = 0$. Therefore, $S \cap S^\perp = \{0\}$. Since the dimensions add up to $\dim(V)$, we conclude that $S \oplus S^\perp = V$. □

Definition 3.26. *The **norm** of a vector is* $\|v\| = \sqrt{\langle v, v \rangle}$. *Two vectors* $v, w \in V$ *are **orthogonal** if* $\langle v, w \rangle = 0$.

Some familiar properties include:

- $\|v - w\|^2 = \langle v - w, v - w \rangle = \|v\|^2 + \|w\|^2 - 2\langle v, w \rangle$.
- If v and w are orthogonal, then $\|v - w\|^2 = \|v\|^2 + \|w\|^2$.
- In general, the angle between two vectors is defined as $\measuredangle(v, w) = \cos^{-1}\left(\dfrac{\langle v, w \rangle}{\|v\| \, \|w\|}\right)$. This definition makes sense only if $|\langle v, w \rangle| \leq \|v\| \, \|w\|$.

Proposition 3.27 (Cauchy-Schwarz Inequality)**.** *For all* $u, v \in V$, $|\langle u, v \rangle| \leq \|u\| \, \|v\|$.

Proof. The inequality is unaffected by scaling, so we assume $\|u\| = 1$. Decompose v along $V = S \oplus S^\perp$, where $S = \operatorname{span}(u) \subset V$. Explicitly, $v = v_1 + v_2$, where $v_1 = \langle v, u \rangle u \in \operatorname{span}(u)$ and $v_2 = v - \langle v, u \rangle u$ is orthogonal to u. Then $v_1 \perp v_2$, so $\|v\|^2 = \|v_1\|^2 + \|v_2\|^2 \geq \|v_1\|^2 = \langle v, u \rangle^2$. This is the desired inequality for $\|u\| = 1$. □

Definition 3.28. *Let* V *be a finite-dimensional vector space over* $\mathbb{R}$ *with inner product* $\langle \cdot, \cdot \rangle$. *A basis* $v_1, \ldots, v_n$ *of* V *is said to be **orthonormal** if*

$$\langle v_i, v_j \rangle = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

In such a basis, $(V, \langle \cdot, \cdot \rangle) \simeq \mathbb{R}^n$ with the standard dot product.

Theorem 3.29. *Every finite-dimensional inner product space over* $\mathbb{R}$ *has an orthonormal basis.*

Two proofs:

Proof. By induction on $\dim(V)$: Choose $v \neq 0 \in V$, let $v_1 = \dfrac{v}{\|v\|}$, so $\|v_1\| = 1$. Let $S = \operatorname{span}(v_1)$, and apply the decomposition $V = S \oplus S^\perp$ (the restriction of $\langle \cdot, \cdot \rangle$ to $S^\perp$ is an inner product). Then $v_1, \ldots, v_n$ is an orthonormal basis for V. □

Proof. Start with any basis $w_1, \ldots, w_n$ of V and apply the Gram-Schmidt process. First set $v_1 = \dfrac{w_1}{\|w_1\|}$. Then take $w_2 - \langle w_2, v_1 \rangle v_1$, which is orthogonal to v_1

(and nonzero by the independence of w_i), and set $v_2 = \dfrac{w_2 - \langle w_2, v_1\rangle v_1}{\|w_2 - \langle w_2, v_1\rangle v_1\|}$, and so on. Set $v_j = \dfrac{w_j - \sum_{i=1}^{j-1}\langle w_j, v_i\rangle v_i}{\|w_j - \sum_{i=1}^{j-1}\langle w_j, v_i\rangle v_i\|}$. Then $(v_1, \ldots, v_n)$ is an orthonormal basis. □

Therefore, every finite-dimensional inner product space over $\mathbb{R}$ is isomorphic (as an inner product space, not just as a vector space) to the standard $\mathbb{R}^n$, where $n = \dim(V)$.

3.5 Orthogonal and Self-Adjoint Operators

Let $(V, \langle\cdot,\cdot\rangle)$ be an inner product space. There are two special classes of linear operators on V that are of particular interest to us.

Definition 3.30. *A linear operator $T : V \to V$ is said to be an **orthogonal operator** if it respects the inner product, i.e.,*

$$\langle Tu, Tv\rangle = \langle u, v\rangle \quad \forall u, v \in V.$$

In other words, *T preserves lengths and angles.*

Remark 3.31.

1. *Orthogonal operators map orthonormal bases to orthonormal bases:*

$$\langle Te_i, Te_j\rangle = \langle e_i, e_j\rangle = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

   In particular, orthogonal operators are always invertible.

2. *If T is orthogonal, then T^{-1} is also orthogonal, since*

$$\langle T^{-1}u, T^{-1}v\rangle = \langle T(T^{-1}u), T(T^{-1}v)\rangle = \langle u, v\rangle \quad \forall u, v.$$

   If T_1 and T_2 are orthogonal, then their product T_1T_2 is also orthogonal. Hence, the set of orthogonal operators forms a subgroup of $Aut(V)$.

3. *If M is the matrix representing T in an orthonormal basis, then $M^TM = I$. To see this, the entries of M^TM are the dot products of the columns of M:*

$$(M^TM)_{ij} = \sum_k M^T_{ik}M_{kj} = \sum_k M_{ki}M_{kj} = \langle M(e_i), M(e_j)\rangle = \langle e_i, e_j\rangle.$$

Definition 3.32. *Let $T : V \to V$ be a linear operator on an inner product space $(V, \langle\cdot,\cdot\rangle)$. There exists a unique linear operator $T^* : V \to V$, called the **adjoint** of T, such that*

$$\langle v, T(w)\rangle = \langle T^*(v), w\rangle \quad \forall v, w \in V.$$

Given $v \in V$, the linear functional $\varphi_v : V \to \mathbb{R}$ defined by $w \mapsto \langle v, T(w)\rangle$ can, by the non-degeneracy of $\langle\cdot,\cdot\rangle$, be written as the inner product of w with some element of V, which we denote $T^*(v)$. Hence, $T^*(v)$ is the unique element of V such that $\langle w, T^*(v)\rangle = \langle v, T(w)\rangle$.

Alternatively, define the isomorphism $\varphi : V \to V^*$ and observe that T^* is the composition

$$V \overset{\varphi}{\to} V^* \overset{T^T}{\to} V^* \overset{\varphi^{-1}}{\to} V,$$

which maps v to $\langle v, \cdot\rangle$, then $\langle v, T(\cdot)\rangle = \langle T^*(v), \cdot\rangle$, yielding $T^*(v)$.

Definition 3.33. *A linear operator $T : V \to V$ is **self-adjoint** if $T^* = T$, i.e.,*

$$\langle T^*(v), w\rangle = \langle T(v), w\rangle \quad \forall v, w \in V.$$

In an orthonormal basis $(e_1, \ldots, e_n)$ of V, if the matrix of T is M, the matrix of T^* is N. We have

$$\langle v, T(w)\rangle = v^T M w, \quad \langle T^*(v), w\rangle = (Nv)^T w = v^T N^T w,$$

which implies $N^T = M$, so $N = M^T$. Hence, in an orthonormal basis, T is self-adjoint if and only if the matrix $M(T)$ is symmetric.

Note that self-adjoint operators need not be invertible. For example, the zero operator is self-adjoint.

Proposition 3.34. *If T is self-adjoint and $S \subset V$ is an invariant subspace of T (i.e., $T(S) \subset S$), then the orthogonal complement $S^\perp$ is also an invariant subspace of T (i.e., $T(S^\perp) \subset S^\perp$).*

Proof. Let $v \in S^\perp$. Then for all $w \in S$, we have $T(w) \in S$, so

$$\langle Tv, w\rangle = \langle v, Tw\rangle = 0$$

(the first equality follows from $T^* = T$, and the second from $v \in S^\perp$ and $T(w) \in S$). Since $\langle Tv, w\rangle = 0$ for all $w \in S$, we conclude that $Tv \in S^\perp$. □

Lemma 3.35. *If T is self-adjoint, then for all $a \in \mathbb{R}_+$, the operator $T^2 + a$ is invertible.*

Proof. For all $v \in V \setminus \{0\}$, we have

$$\langle (T^2 + a)v, v\rangle = \langle T^2 v, v\rangle + a\langle v, v\rangle = \langle Tv, Tv\rangle + a\langle v, v\rangle = \|Tv\|^2 + a\|v\|^2 \geq 0.$$

Thus, $(T^2 + a)v \neq 0$, implying that $\mathrm{Ker}(T^2 + a) = \{0\}$. □

Corollary 3.36. *If $p(x) \in \mathbb{R}[x]$ is a quadratic polynomial with no real roots and $T^* = T$, then $p(T)$ is invertible.*

Proof. It suffices to show that $T^2 + bT + c$ is invertible whenever $b^2 - 4c < 0$. We can write

$$T^2 + bT + c = \left(T + \frac{b}{2}\right)^2 + a, \quad a = c - \frac{b^2}{4} > 0,$$

and since $T + \frac{b}{2}$ is self-adjoint, by the lemma, $T^2 + bT + c$ is invertible. □

Proposition 3.37 (The Spectral Theorem for Real Self-Adjoint Operators)**.** *If $T : V \to V$ is self-adjoint, then T is diagonalizable with real eigenvalues. Moreover, T can be diagonalized in an orthonormal basis of $(V, \langle \cdot, \cdot \rangle)$.*

Proof. First, we show the existence of an eigenvector. Pick $v \in V$, $v \neq 0$. Since the vectors $v, Tv, T^2v, \ldots, T^nv$ are linearly dependent, there exists a non-constant polynomial such that

$$(a_nT^n + \cdots + a_0)v = 0.$$

This polynomial factors into linear and quadratic factors over $\mathbb{R}$:

$$\prod (T - \lambda_i) \prod (T^2 + b_jT + c_j)v = 0,$$

where the first product corresponds to real eigenvalues and the second to irreducible quadratics corresponding to complex conjugate eigenvalues.

At least one of these operators must have a nontrivial kernel (otherwise their product would be invertible, which would imply $v = 0$). By the previous corollary, each $T^2 + b_jT + c_j$ is invertible, so some $T - \lambda_i$ must have a nontrivial kernel, thus yielding an eigenvector!

Now, diagonalization: we know there is an eigenvector $v_1 \in V$ with eigenvalue $\lambda_1 \in \mathbb{R}$. Scaling v_1 if needed, we may assume $\|v_1\| = 1$. The subspace $S = \text{span}(v_1) \subset V$ is invariant under T, and by the Spectral Theorem, $S^\perp$ is also invariant under T. By induction, restricting $\langle \cdot, \cdot \rangle$ to $S^\perp$, we find that there is an orthonormal basis $(v_2, \ldots, v_n)$ of eigenvectors of T, and $(v_1, \ldots, v_n)$ forms a basis of V in which T is diagonal. □

Corollary 3.38. *If T is self-adjoint, then the matrix of T in a suitable orthonormal basis is diagonal:*

$$M(T) = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

Remark 3.39. *This also implies that eigenvectors of T corresponding to distinct eigenvalues are orthogonal. This can be seen as follows: if $Tv = \lambda v$ and $Tw = \mu w$, then*

$$\lambda \langle v, w \rangle = \langle Tv, w \rangle = \langle v, Tw \rangle = \mu \langle v, w \rangle.$$

Thus, if $\lambda \neq \mu$, we must have $\langle v, w \rangle = 0$, so $v \perp w$.

Back to orthogonal transformations: Do we have a similar structure/result?

- In dimension 1: T is multiplication by a scalar, so T orthogonal $\iff T = \pm I$.
- In dimension 2: T orthogonal $\iff T$ is a rotation or reflection. Given an orthonormal basis (e_1, e_2), Te_1 is any unit vector on the unit circle, $\{v \in V \mid \|v\| = 1\} = \{\cos\theta e_1 + \sin\theta e_2\}$, and Te_2 is also a unit vector orthogonal to Te_1, implying two possibilities: the rotation matrix or the reflection matrix.

The rotation matrix by θ degrees is

$$\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

which has no eigenvectors.

The reflection matrix is

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix},$$

which has eigenvalues ± 1 and two orthogonal eigenspaces.

Notation for $(V, \langle\cdot,\cdot\rangle)$: The subgroups $SO(V) \subset O(V) \subset GL(V)$ are defined as follows:

- $SO(V)$ is the subgroup of orientation-preserving orthogonal transformations, those with $\det = \pm 1$. In dimension 1: $\{\pm I\}$. In dimension 2: rotations.
- $O(V)$ is the orthogonal group of V.
- $GL(V)$ is the group of invertible linear operators $T : V \to U$.

Since $V \simeq \mathbb{R}^d$ by choosing an orthonormal basis, we typically write $O(\mathbb{R}^n) = O(n)$ and $SO(\mathbb{R}^n) = SO(n)$. We also have the short exact sequence:

$$1 \to SO(n) \to O(n) \to \{\pm 1\} = \mathbb{Z}/2 \to 1,$$

where $SO(n)$ has index 2 in $O(n)$, and $SO(2) \simeq S^1$ (rotations correspond to angles).

Proposition 3.40. *If $T : V \to V$ is an orthogonal operator on a finite-dimensional inner product space, then V decomposes into a direct sum of orthogonal invariant subspaces $V = \bigoplus V_i$, where $V_i \perp V_j$ for $i \neq j$ and $T(V_i) = V_i$, with $\dim V_i = 1$ or 2. Specifically:*

- *If $\dim V_i = 1$, then $T|_{V_i} = \pm I$.*
- *If $\dim V_i = 2$, then $T|_{V_i}$ is either a rotation or a reflection.*

In the latter case, we can further decompose into ± 1 eigenspaces, so we can replace reflections by 1-dimensional blocks.

This provides a nice way to think about an individual transformation as built from reflections and rotations on individual subspaces, but it doesn't help in understanding the composition of two orthogonal transformations (whose invariant subspaces may not coincide). For example, in $\mathbb{R}^3$, is there a nice formula for the product of two rotations?

3.6 Hermitian Inner Products

Now, let's examine the analogue of inner products for complex vector spaces: Hermitian inner products. As noted previously, a bilinear form on a complex vector space $V \times V \to \mathbb{C}$ cannot be positive-definite, since $b(iv, iv) = -b(v, v)$. The solution is to abandon $\mathbb{C}$-linearity in one of the two variables and instead require "conjugate linearity."

Definition 3.41. *A **Hermitian form** on a complex vector space V is a map $H : V \times V \to \mathbb{C}$ such that H is **sesquilinear**, i.e.,*

- $H(u + v, w) = H(u, w) + H(v, w)$ *and* $H(u, v + w) = H(u, v) + H(u, w)$,
- $H(u, \lambda v) = \lambda H(u, v)$, *but* $H(\lambda u, v) = \overline{\lambda} H(u, v)$,

*and H is **conjugate symmetric**, i.e., $H(u, v) = \overline{H(v, u)}$.*

Conjugate symmetry implies that $H(u, u) \in \mathbb{R}$ for all $u \in V$.

Definition 3.42. *A **Hermitian inner product** is a positive-definite (conjugate symmetric) Hermitian form.*

Remark 3.43. *The map*

$$\varphi_u : V \to V^*, \quad u \mapsto H(u, \cdot),$$

*is now a **complex antilinear** map, meaning that $\varphi(\lambda u) = \overline{\lambda}\varphi(u)$ for all $\lambda \in \mathbb{C}$.*

Various properties carry over from the real case:

- If H is positive-definite, then H is non-degenerate (i.e., $\mathrm{Ker}(\varphi_H) = 0$),
- Given a subspace $W \subset V$, its orthogonal complement $W^\perp = \{v \in V \mid H(v, w) = 0 \, \forall w \in W\}$ is also a subspace, and $V = W \oplus W^\perp$. Conjugate linearity does not affect the fact that $W^\perp$ is a $\mathbb{C}$-subspace. Moreover, positivity implies that $W \cap W^\perp = \{0\}$.

The following property also holds:

Definition 3.44. *An **orthonormal basis** of V with a Hermitian inner product is a basis $\{e_i\}$ such that*

$$H(e_i, e_j) = \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Proposition 3.45. *Every finite-dimensional Hermitian inner product space admits an orthonormal basis.*

This can be proven similarly to the real case by induction on $\dim(V)$. Start by choosing a vector v_1 such that $||v_1||^2 = H(v_1, v_1) = 1$, then take an orthonormal basis $\{v_2, \ldots, v_n\}$ of the orthogonal complement of $\text{span}(v_1)$, or apply the Gram-Schmidt process.

Corollary 3.46. *Every finite-dimensional Hermitian inner product space is isomorphic to $\mathbb{C}^n$ with the standard Hermitian inner product,*

$$H(z, w) = \sum_j \overline{z_j} w_j.$$

In matrix form, this is written as

$$H(z, w) = z^* w,$$

where $z^* = \overline{z}^T = (\overline{z_1}, \ldots, \overline{z_n})$ is the conjugate transpose.

Example 3.47 (Fourier Series)**.** *This is a not-quite-example. Let $V = C^\infty(S^1, \mathbb{C})$, the space of infinitely differentiable functions from $S^1 \simeq \mathbb{R}/\mathbb{Z}$ to $\mathbb{C}$, and define*

$$\langle f, g \rangle = \int_{S^1} \overline{f(t)} g(t)\, dt.$$

The functions $f_n(t) = e^{2\pi i n t}$ form an orthogonal set, with $\langle f_n, f_m \rangle = \delta_{mn}$. The set $\{f_n\}_{n \in \mathbb{Z}}$ is not a basis for V, but their span $W \subset V$ is the space of trigonometric polynomials. We can think of the Fourier series as an orthogonal projection onto W. This becomes clearer with analysis or, even better, with the theory of Hilbert spaces.

Definition 3.48. *Let V be a complex vector space, H a Hermitian inner product, and $T : V \to V$ a linear map. The following definitions hold:*

- *The **adjoint** of T is $T^* : V \to V$ such that*

$$H(T^* v, w) = H(v, Tw) \quad \forall v, w \in V.$$

- *T is **self-adjoint** if $T^* = T$, i.e., $H(v, Tw) = H(Tv, w)$ for all $v, w \in V$.*
- *T is **unitary** if $H(Tv, Tw) = H(v, w)$ for all $v, w \in V$, i.e., $T^* = T^{-1}$.*

Unitary operators form a subgroup $U(V, H) \subset \text{Aut}(V)$ (and $U(n) \subset GL(n, \mathbb{C})$). Note that $U(1) \simeq S^1$ (the group of complex numbers with norm 1, i.e., multiplication by any complex number of norm 1).

In an orthonormal basis, the matrix of the adjoint $M(T^*)$ is the conjugate transpose of the matrix of T:

$$M(T^*) = M(T)^* = \overline{M(T)^T}.$$

This follows from the identity

$$H(Tv, w) = (Mv)^* w = v^* M^* w = H(v, T^* w).$$

Thus, self-adjoint complex operators are represented by Hermitian matrices: $a_{ij} = \overline{a_{ji}}$.

Now we state the complex spectral theorem.

Proposition 3.49 (The Complex Spectral Theorem)**.** *If V is a finite-dimensional complex vector space, $H : V \times V \to \mathbb{C}$ is a Hermitian inner product, and $T : V \to V$ is self-adjoint (i.e., $T^* = T$) or unitary (i.e., $T^* = T^{-1}$), then there exists an orthonormal basis consisting of eigenvectors of T. Thus, T is diagonalizable, with eigenvalues in $\mathbb{R}$ if self-adjoint, or in the unit circle S^1 if unitary.*

Proof. As in the real case, the key observation is that if $S \subset V$ is invariant under T (i.e., $T(S) \subset S$), then so is $S^\perp \subset V$. In both cases, if S is invariant for T, it is also invariant for $T^* = T^{\pm 1}$. Thus, if $v \in S^\perp$, for all $w \in S$, we have

$$H(Tw, w) = H(v, T^* w) = 0.$$

Starting with an eigenvector v_1 such that $Tv_1 = \lambda_1 v_1$ and $||v_1|| = 1$, we let $S = \text{span}(v_1)$ and consider the restriction of T to $S^\perp$. □

Returning to (non-degenerate) symmetric bilinear forms:

Suppose V is a finite-dimensional vector space over a field k and $B : V \times V \to k$ is a non-degenerate symmetric bilinear form. Can we classify such forms?

Remark 3.50. *Define the quadratic form $Q(v) = B(v, v) : V \to k$. Note that Q is not necessarily positive-definite unless B is positive-definite. However, if $k = \mathbb{R}$, we can still classify these forms using the Spectral Theorem.*

Classification approach: Find a vector v such that $B(v, v) \neq 0$, and then consider the orthogonal complement of the span of v, denoted $\text{span}(v)^\perp$. (Note that $\text{span}(v)^\perp = \text{Ker}(\varphi_B(v))$, where $\varphi_B(v) : v \mapsto k$.) Thus, when $B(v, v) \neq 0$, we have the direct sum decomposition $V = \text{span}(v) \oplus \text{span}(v)^\perp$. Next, study the restriction of B to $\text{span}(v)^\perp$, i.e., $B_{|\text{span}(v)^\perp}$, and so on.

Proposition 3.51. *Over $\mathbb{C}$, any nondegenerate symmetric bilinear form admits a basis $\{e_1, \ldots, e_n\}$ such that $B(e_i, e_j) = \delta_{ij}$.*

Remark 3.52. *Hermitian forms are typically of greater interest in many applications, however.*

Proof. Since $B(u, v) \neq 0$, it follows that at least one of $B(u, u)$, $B(v, v)$, or $B(u + v, u + v)$ must be nonzero. Therefore, the nondegeneracy of B implies the existence of a vector v such that $B(v, v) \neq 0$. Let $e_1 = B(v, v)^{-\frac{1}{2}} v$. Consider the orthogonal complement of $\text{span}(e_1)$, denoted $W = \text{span}(e_1)^\perp$.

Additionally, since $B(e_1, e_1) \neq 0$, we have $\text{span}(e_1) \cap \text{span}(e_1)^\perp = \{0\}$, and $\dim(W) = \dim(\text{Ker}(B(e_1, \cdot))) = \dim(V) - 1$. Hence, we can express V as a direct sum: $V = \text{span}(e_1) \oplus W$.

The restriction of B to W is nondegenerate because the matrix of B in the basis $\{e_1, \text{some basis of } W\}$ is

$$\left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & B_{|W} \end{array}\right)$$

which is invertible (rank n) if and only if $B_{|W}$ is invertible (rank $n-1$). We can complete the proof by induction on the dimension. Assuming the result holds for $\dim(V) = n-1$, we can extend it to $\dim(V) = n$ by choosing an appropriate basis for V such that $B_{|W}(e_j, e_k) = \delta_{jk}$ for all j, k. □

Proposition 3.53. *Over $\mathbb{R}$, any nondegenerate symmetric bilinear form admits a basis such that*

$$B(e_i, e_j) = \begin{cases} 0 & \text{if } i \neq j, \\ \pm 1 & \text{if } i = j. \end{cases}$$

That is, we can assume that for any linear combination $\sum_{i=1}^n x_i e_i$ and $\sum_{i=1}^n y_i e_i$, the bilinear form satisfies

$$B\left(\sum_{i=1}^n x_i e_i, \sum_{i=1}^n y_i e_i\right) = \sum_{i=1}^k x_i y_i - \sum_{k+1}^n x_i y_i,$$

where

$$B = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & -1 & & \\ & & & & \ddots & \\ & & & & & -1 \end{pmatrix}$$

We say that a bilinear form B has **signature** $(k, n-k)$, where the case $(n, 0)$ corresponds to a definite positive form. Here, k is the maximum dimension of a subspace $W \subseteq V$ such that the restriction $B_{|W}$ is definite positive, and $n-k$ is the maximum dimension of a subspace $W \subseteq V$ such that $B_{|W}$ is definite negative.

Proof. The proof is the same as in the complex case, except that in the real case we cannot always scale to $B(e_1, e_1) = 1$. Instead, we can only force $B(e_1, e_1) = \pm 1$. □

Over $\mathbb{Q}$, the situation becomes much more complicated—number theory enters the picture!

Example 3.54. *Consider the bilinear form* $B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. *The associated quadratic form is* $Q(v) = B(v,v) = v_1^2 + v_2^2$. *However, there does not exist a vector* $v = (v_1, v_2) \in \mathbb{Q}^2$ *such that* $B(v,v) = v_1^2 + v_2^2 = 3$.

To see this, suppose there are integers $n_1, n_2, m \in \mathbb{Z}$ *with no common factor such that* $n_1^2 + n_2^2 = 3m^2$. *Since* $n_1^2 + n_2^2 \equiv 0, 1, 2 \pmod 4$ *and* $3m^2 \equiv 0, 3 \pmod 4$, *it follows that* $n_1^2 + n_2^2 \equiv 3m^2 \pmod 4$, *which leads to a contradiction, as both must be congruent to* 0 *or* $1 \pmod 4$.

In contrast, the bilinear form $B' = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ *does admit a solution. Specifically, there exists a vector* $v = (1,1)$ *such that* $B'(v,v) = 3$.

Now, consider the skew-symmetric case. Suppose $\mathrm{char}(k) \neq 2$. We can still find a standard basis for a finite-dimensional vector space V with a nondegenerate skew-symmetric bilinear form $B : V \times V \to k$ (also known as a symplectic form), but the process is slightly different since $B(v,v) = 0$ for all $v \in V$.

To begin, pick any nonzero $e_1 \in V$. Since V is nondegenerate, the map $B(e_1, \cdot) : V \to k$ is nonzero, implying the existence of a vector $f_1 \in V$ such that $B(e_1, f_1) \neq 0$. We can scale f_1 so that $B(e_1, f_1) = 1$. Now consider the subspace $\mathrm{span}(e_1, f_1)$. Since B is skew-symmetric, we have $\mathrm{span}(e_1, f_1) \cap \mathrm{span}(e_1, f_1)^\perp = \{0\}$, because if $v = ae_1 + bf_1$ satisfies $B(v, e_1) = 0$ and $B(v, f_1) = 0$, it follows that $a = b = 0$. Therefore, we can write V as the direct sum $V = \mathrm{span}(e_1, f_1) \oplus \mathrm{span}(e_1, f_1)^\perp$. We then study the restriction of B to the subspace $\mathrm{span}(e_1, f_1)^\perp$ using induction on the dimension. This leads to the following result:

Proposition 3.55. *Let* V *be a finite-dimensional vector space over a field* k *with* $\mathrm{char}(k) \neq 2$, *and let* B *be a nondegenerate skew-symmetric bilinear form on* V. *Then the dimension of* V *is even, and there exists a basis* $\{e_1, f_1, \ldots, e_n, f_n\}$ *of* V *such that:*

$$B(e_i, e_j) = B(f_i, f_j) = 0, \quad B(e_i, f_j) = \delta_{ij} = -B(f_j, e_i).$$

That is, the matrix of B in this basis is

$$\begin{pmatrix} 0 & 1 & & & \\ -1 & 0 & & & \\ & & 0 & 1 & \\ & & -1 & 0 & \\ & & & & \ddots \end{pmatrix}$$

The group of linear transformations preserving B is the **symplectic group** $Sp(V, B) \simeq Sp(2n, k)$.

3.7 Tensor Products: Definition and Basic Properties

Let V, W be finite-dimensional vector spaces over k. The **tensor product** is a vector space $V \otimes W$ with a bilinear map

$$V \times W \to V \otimes W$$
$$(v, w) \mapsto v \otimes w.$$

There are three equivalent definitions (from concrete to abstract; all give the same output up to natural isomorphism).

Definition 3.56. *Choose bases $e_1, \ldots, e_m$ of V, and $f_1, \ldots, f_n$ of W. Then $V \otimes W$ is the vector space with basis $\{e_i \otimes f_j \mid 1 \leq i \leq m, 1 \leq j \leq n\}$.*

The bilinear map is $(e_i, f_j) \mapsto e_i \otimes f_j$, and it extends by linearity.

Elements of the form $v \otimes w = (\sum a_i e_i) \otimes (\sum b_j f_j) = \sum a_i b_j (e_i \otimes f_j)$ are called **pure tensors**. Not every element of $V \otimes W$ is of this form! The **rank** of an element of $V \otimes W$ is the minimum number of terms needed to express it as a linear combination of pure tensors.

This definition is concrete and shows that $\dim(V \otimes W) = mn$, but the independence of the choice of basis isn't obvious. To de-emphasize the basis:

Definition 3.57. *Start with a vector space U with basis $\{v \otimes w \mid v \in V, w \in W\}$ (this is often an uncountably large basis), and quotient it by a subspace R of relations among these elements. Specifically, R is the span of the following relations:*

$$(\lambda v) \otimes w - \lambda(v \otimes w),$$
$$v \otimes (\lambda w) - \lambda(v \otimes w),$$
$$(u + v) \otimes w - u \otimes w - v \otimes w,$$
$$u \otimes (v + w) - u \otimes v - u \otimes w,$$

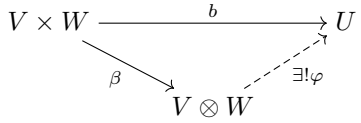
for all scalars λ and vectors u, v, w.

Defining $V \otimes W = U/R$ sets all these relations to zero, enforcing the bilinearity of the map $(v, w) \mapsto v \otimes w$.

This definition shows the independence of the choice of basis but involves a large construction. Ultimately, if we have bases $\{e_i\}$ of V and $\{f_j\}$ of W, the relations in R ensure that all elements of $V \otimes W$ are linear combinations of $e_i \otimes f_j$. However, before checking this, it's not even obvious that $\dim(V \otimes W) < \infty$.

The least concrete, yet most mathematically satisfactory, definition characterizes what $V \otimes W$ does without specifying its construction. Namely, $V \otimes W$ is the largest space through which all bilinear maps from $V \times W$ factor. (For example, in the second definition, U is too large, and quotienting by R enforces bilinearity.)

Definition 3.58. *The **tensor product** $V \otimes W$ is the universal vector space through which all bilinear maps from $V \times W$ factor. That is, it is a vector space $V \otimes W$ and a bilinear map $\beta : V \times W \to V \otimes W$ such that, given any vector space U over k, and any bilinear map $b : V \times W \to U$, there exists a unique linear map $\varphi : V \otimes W \to U$ such that $b = \varphi \circ \beta$:*



This definition tells us the key property of $V \otimes W$ and implies uniqueness up to isomorphism (the universal property gives isomorphisms between any two candidate constructions of $V \otimes W$). Existence ultimately follows from one of the previous constructions.

Check: definition 1 satisfies the property. Given bases $\{e_i\}$ and $\{f_j\}$ of V and W, we have

$$\{\text{bilinear maps } b : V \times W \to U\} \leftrightarrow \{\text{linear maps } \varphi : V \otimes W \to U\},$$

by defining $b(e_i, f_j) = \varphi(e_i \otimes f_j)$ and vice versa.

Now, let's move on to some basic properties:

Proposition 3.59. *$\otimes : Vect_k \times Vect_k \to Vect_k$ is a functor.*

This means that, given linear maps $f : V \to V'$, $g : W \to W'$, we get a linear map $f \otimes g : V \otimes W \to V' \otimes W'$ on pure elements: $(f \otimes g)(v \otimes w) = f(v) \otimes g(w)$, and this respects composition.

Proposition 3.60.

$$V \otimes W \cong W \otimes V.$$

This is a natural isomorphism, and we could even claim that they are equal.

Proposition 3.61.

$$(U \oplus V) \otimes W \cong (U \otimes W) \oplus (V \otimes W).$$

This is more surprising but extremely useful:

Proposition 3.62.

$$Hom(V, W) \simeq V^* \otimes W.$$

Proof. The map

$$\begin{aligned} V^* \times W &\to \mathrm{Hom}(V, W) \\ (\ell, w) &\mapsto (v \mapsto \ell(v)w) \end{aligned}$$

is bilinear. So, by the universal property, we get a linear map $V^* \otimes W \to \mathrm{Hom}(V, W)$ that takes $\ell \otimes w \mapsto (v \mapsto \ell(v)w)$. Pick bases $(e_1, \ldots, e_n)$ of V, $(f_1, \ldots, f_m)$ of W, and let $(e_1^*, \ldots, e_n^*)$ be the dual basis of V^*. Then $(e_i^* \otimes f_j)$ is a basis of $V^* \otimes W$. The above construction takes $(e_i^* \otimes f_j)$ to

$$\begin{aligned} \varphi_{ij} : V &\to W \\ v &\mapsto e_i^*(v) f_j, \end{aligned}$$

whose action on basis elements is that e_i maps to f_j, and all others map to 0. Thus, $M(\varphi_{ij})$ is an $m \times n$ matrix with a single nonzero entry in the ith column and jth row. These form a basis of $\mathrm{Hom}(V, W)$. Since it maps a basis to a basis, $V^* \otimes W \to \mathrm{Hom}(V, W)$ is an isomorphism. □

Example 3.63. *If V has a basis (e_1, e_2), V^* has a basis (e_1^*, e_2^*), and W has a basis (f_1, f_2), then the linear map with matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is*

$$e_1^* \otimes (af_1 + cf_2) + e_2^* \otimes (bf_1 + df_2).$$

This is generally a rank 2 tensor, except if $ad - bc = 0$, in which case we can write it as a pure tensor

$$(xe_1^* + ye_2^*) \otimes (zf_1 + wf_2).$$

Proposition 3.64. *The tensor rank in $V^* \otimes W$ is the same as the rank in $Hom(V, W)$ (hence the name).*

For the rank 1 case: $\ell \otimes w$ corresponds to $(v \mapsto \ell(v)w)$, whose image is $\mathrm{span}(w)$. This is easiest to see if we take a basis of V in which $e_{r+1}, \ldots, e_n$ is a basis of $\mathrm{Ker}(\varphi)$ and of W in which $f_1, \ldots, f_r$ is a basis of $\mathrm{Im}(\varphi)$, with $f_i = \varphi(e_i)$ for all $1 \le i \le r$. Then φ corresponds to $\sum_{i=1}^r e_i^* \otimes f_i$, and the matrix of φ is

$$\left(\begin{array}{ccc|c} 1 & & & \\ & \ddots & & 0 \\ & & 1 & \\ \hline & 0 & & 0 \end{array}\right)$$

where the width of the matrix is $r = \mathrm{rank}(\varphi)$.

The isomorphism $\mathrm{Hom}(V, W) \simeq V^* \otimes W$ also implies:

- $(V \otimes W)^* \simeq V^* \otimes W^*$. This can be seen as follows:

$$\begin{aligned} (V \otimes W)^* &= \mathrm{Hom}(V \otimes W, k) \\ &= \{\text{Bilinear maps } V \times W \to k\} \\ &\simeq \mathrm{Hom}(V, W^*) \\ &\simeq V^* \otimes W^*. \end{aligned}$$

  The first $\simeq$ is given by $b \mapsto \varphi_b$, where $\varphi_b(v) = b(v, \cdot)$, with $v \in V$ and $b(v, \cdot) \in W^*$.

- $\mathrm{Hom}(V, W) \simeq V^* \otimes W = (W^*)^* \otimes V^* \simeq \mathrm{Hom}(W^*, V^*)$. This is the transpose construction, where $\varphi \in \mathrm{Hom}(V, W)$ corresponds to $\varphi^t : W^* \to V^*$.

We can now properly define the trace of a linear operator! In standard linear algebra classes, the trace of an $n \times n$ matrix $A = (a_{ij})$ is defined as $\mathrm{tr}(A) = \sum_{i=1}^n a_{ii}$, the sum of the diagonal entries. Noting that $\mathrm{tr}(AB) = \sum_{i,j} a_{ij} b_{ji} = \mathrm{tr}(BA)$, we have $\mathrm{tr}(P^{-1}AP) = \mathrm{tr}(A)$, and so the trace of $T : V \to V$ is defined as the trace of $\mathcal{M}(T)$ in any basis. We could also define the trace via eigenvalues and their multiplicities, and for an algebraically closed field, in a basis where $\mathcal{M}(T)$ is triangular, we see that $\mathrm{tr}(T) = \sum n_i \lambda_i$.

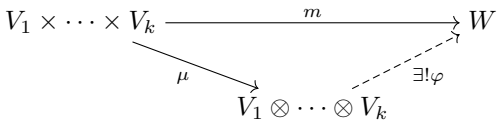
We can improve this definition (conceptually) by using $\mathrm{Hom}(V, V) \simeq V^* \otimes V$, and the contraction linear map $V^* \otimes V \to k$. Namely, there's a natural bilinear pairing

$$\begin{aligned} \mathrm{ev} : V^* \times V &\to k \\ (\ell, v) &\mapsto \ell(v) \end{aligned}$$

which determines $\mathrm{tr} : V^* \otimes V \to k$. On pure tensors, $\ell \otimes v \mapsto \ell(v)$. This is indeed equivalent to the usual definition: choosing a basis (e_i) and the dual basis (e_i^*), $\mathrm{tr}(e_i^* \otimes e_j) = e_i^*(e_j) = \delta_{ij}$, which corresponds to the trace of the matrix with a single entry 1 in position (j, i).

Definition 3.65. *A map $m : V_1 \times \cdots \times V_k \to W$ is called **multilinear** if it is linear in each variable separately.*

The tensor product $V_1 \otimes \cdots \otimes V_k$ can be defined in various ways: either by using bases of $V_1, \ldots, V_k$, or as a quotient of a universal vector space by certain relations, or through the universal property for multilinear maps. Specifically, there exists a multilinear map $\mu : V_1 \times \cdots \times V_k \to V_1 \otimes \cdots \otimes V_k$ such that $(v_1, \ldots, v_k) \mapsto v_1 \otimes \cdots \otimes v_k$. Moreover, for any vector space W, and for any multilinear map $m : V_1 \times \cdots \times V_k \to W$, there exists a unique map $\varphi \in \mathrm{Hom}(V_1 \otimes \cdots \otimes V_k, W)$ such that $m = \varphi \circ \mu$:



In fact, nothing new is happening here, since we have the well-known isomorphism $(U \otimes V) \otimes W = U \otimes (V \otimes W) = U \otimes V \otimes W$. However, in the special case where $V^{\otimes n} = V \otimes \cdots \otimes V$ (with n copies of V, and by convention $V^{\otimes 0} = k$, $V^{\otimes 1} = V$), we obtain bilinear maps $V^{\otimes k} \times V^{\otimes \ell} \to V^{\otimes (k+\ell)}$ for all $k, \ell \geq 0$. Taken together, these maps define a multiplication on the **tensor algebra**

$$T(V) = \bigoplus_{n=0}^{\infty} V^{\otimes n},$$

making it a noncommutative ring.

3.8 Symmetric and Exterior Algebra

Recall that the space of bilinear forms $B(V) \simeq V^* \otimes V^*$ decomposes into

$$B(V) = B_{\text{sym}} \oplus B_{\text{skew}}$$

where B_{sym} and B_{skew} represent the symmetric and skew-symmetric bilinear forms, respectively. Equivalently, there is an involution (an automorphism such that $\varphi^2 = \text{id}$) $\varphi : B(V) \to B(V)$ that maps $b(x, y) \mapsto b(y, x)$, or equivalently, on $V^* \otimes V^*$: $\ell \otimes \ell' \mapsto \ell' \otimes \ell$.

This involution φ has eigenvalues ± 1, with:

$$\text{Ker}(\varphi - I) = B_{\text{sym}}, \quad \text{Ker}(\varphi + I) = B_{\text{skew}}.$$

We can also extend this to higher tensor powers of V or V^* (in the latter case, considering multilinear forms).

There is an action of the symmetric group S_d on $V^{\otimes d}$, i.e., each permutation $\sigma \in S_d$ defines a linear map

$$V^{\otimes d} \xrightarrow{\sigma} V^{\otimes d} \quad \text{with} \quad v_1 \otimes \cdots \otimes v_d \mapsto v_{\sigma(1)} \otimes \cdots \otimes v_{\sigma(d)}.$$

This defines a group homomorphism $S_d \to \text{Aut}(V^{\otimes d})$.

Definition 3.66. *A tensor $\eta \in V^{\otimes d}$ is **symmetric** if $\sigma \cdot \eta = \eta$ for all $\sigma \in S_d$. The space of symmetric tensors is denoted $\text{Sym}^d(V) \subset V^{\otimes d}$.*

For example, $\text{Sym}^d(V^*)$ consists of symmetric multilinear forms $m : V \times \cdots \times V \to k$, satisfying $m(v_{\sigma(1)}, \ldots, v_{\sigma(d)}) = m(v_1, \ldots, v_d)$.

If $\text{char}(k) = 0$, the symmetric part of a tensor can be obtained by averaging:

$$\alpha : V^{\otimes d} \to \text{Sym}^d(V), \quad \alpha(v_1 \otimes \cdots \otimes v_d) = \frac{1}{d!} \sum_{\sigma \in S_d} v_{\sigma(1)} \otimes \cdots \otimes v_{\sigma(d)}.$$

Definition 3.67. *If $\text{char}(k) = 0$, we can also define $\text{Sym}^d(V)$ as the quotient of $V^{\otimes d}$ by the subspace spanned by elements of the form $\eta - \sigma(\eta)$, for $\sigma \in S_d$. Explicitly, this quotient is generated by elements like*

$$v_1 \otimes v_2 \otimes v_3 \otimes \cdots \otimes v_d - v_2 \otimes v_1 \otimes v_3 \otimes \cdots \otimes v_d,$$

where transpositions generate S_d, and similarly for swapping other factors. This definition is different from, but isomorphic to, the previous one.

To determine which definition (as a quotient versus subspace of $V^{\otimes d}$) is preferable, we use a universal property.

Recall that $V^{\otimes d}$ comes with a multilinear map $\mu : V^d \to V^{\otimes d}$, and it is characterized by the isomorphism:

$$\text{Hom}(V^{\otimes d}, U) \simeq \{\text{multilinear maps } V^d \to U\}, \quad \varphi \mapsto \varphi \circ \mu.$$

Now, $\mathrm{Sym}^d V$ comes with a symmetric multilinear map $V^d \to \mathrm{Sym}^d V$, and it is characterized by:

$$\mathrm{Hom}(\mathrm{Sym}^d V, U) \simeq \{\text{symmetric multilinear maps } V^d \to U\}.$$

Definition 3.68. *The product operations $V^{\otimes k} \times V^{\otimes \ell} \to V^{\otimes k+\ell}$ induce a product $Sym^k V \times Sym^\ell V \to Sym^{k+\ell} V$ (using $\otimes$ followed by averaging with α). These combine into a product operation on the symmetric algebra $Sym^\cdot(V) := \bigoplus_{d \geq 0} Sym^d(V)$, called the **symmetric algebra** of V.*

Proposition 3.69. *The symmetric algebra $Sym^\cdot(V)$ is a commutative ring.*

Proof. The product remains associative despite the symmetrization by averaging:

$$\alpha(\alpha(u \otimes v) \otimes w) = \alpha(u \otimes \alpha(v \otimes w)) = \alpha(v \otimes v \otimes w).$$

□

Concretely, if $e_1, \ldots, e_n$ is a basis of V, then

$$\mathrm{Sym}^\cdot(V) \simeq k[e_1, \ldots, e_n]$$

is the algebra of polynomial expressions in the formal variables $e_1, \ldots, e_n$.

This can be seen by denoting $\alpha(e_{i_1} \otimes \cdots \otimes e_{i_k})$ by $e_{i_1} \cdots e_{i_k}$, and considering finite linear combinations of all such terms.

More explicitly: if $e_1, \ldots, e_n$ is a basis of V, then any linear form on V, $\ell \in V^*$, is of the form $v = \sum x_i e_i \mapsto \ell(v) = \sum a_i x_i$, which is a degree 1 polynomial.

Symmetric multilinear forms $\eta \in \mathrm{Sym}^d V^*$ are likewise polynomials (with only degree d terms):

$$v = \sum x_i e_i \mapsto \eta(v, \ldots, v) = \sum_{i_1, \ldots, i_d} x_{i_1} \cdots x_{i_d}.$$

Thus, we have the following result:

Proposition 3.70.

$$Sym^\cdot(V^*) \simeq k[x_1, \ldots, x_n]$$

where the x_i are considered as the coordinates of a vector in V, viewed as linear (degree 1) polynomials on V (i.e., this is another notation for $e_i^* \in V^*$).

Next, we perform a similar construction for skew-symmetric (alternating) multilinear forms.

Definition 3.71. *A tensor $\eta \in V^{\otimes d}$ is **alternating** if $\sigma(\eta) = (-1)^\sigma \eta$ for all $\sigma \in S_d$, where $(-1)^\sigma$ is the sign of σ: -1 for transpositions and products of an odd number of them. The space of alternating tensors is denoted $\Lambda^d(V) \subset V^{\otimes d}$.*

In characteristic zero, we can view $\Lambda^d(V)$ as the image of the skew-symmetrization operator:

$$\beta : V^{\otimes d} \to V^{\otimes d}, \quad v_1 \wedge \cdots \wedge v_d := \beta(v_1 \otimes \cdots \otimes v_d) = \frac{1}{d!} \sum_{\sigma \in S_d} (-1)^\sigma v_{\sigma(1)} \otimes \cdots \otimes v_{\sigma(d)}.$$

This is zero whenever $v_i = v_j$ for some $i \neq j$, and by multilinearity, it is also zero whenever $v_1, \ldots, v_d$ are linearly dependent. Thus, we have $\Lambda^d(V) = 0$ whenever $d > \dim V$.

Alternative definitions:

Definition 3.72. *$\Lambda^d(V)$ can be defined as the quotient of $V^{\otimes d}$ by the subspace spanned by elements like:*

$$v_1 \otimes v_2 \otimes \cdots \otimes v_d + v_2 \otimes v_1 \otimes \cdots \otimes v_d,$$

and similarly for other transpositions swapping two factors.

Or alternatively:

Definition 3.73. *$\Lambda^d(V)$ can be defined as the vector space with an alternating multilinear map*

$$V \times \cdots \times V \to \Lambda^d V, \quad (v_1, \ldots, v_d) \mapsto v_1 \wedge \cdots \wedge v_d,$$

where $v_1 \wedge v_2 = -v_2 \wedge v_1$, and so on. This is universal for alternating multilinear maps on $V \times \cdots \times V$.

Proposition 3.74. *If $(e_1, \ldots, e_n)$ is a basis of V, then the set $\{e_{i_1} \wedge \cdots \wedge e_{i_d} \mid i_1 < \cdots < i_d\}$ forms a basis of $\Lambda^d V$.*

There is a product $\Lambda^k V \times \Lambda^\ell V \to \Lambda^{k+\ell} V$ induced by tensor algebra and skew-symmetrization:

$$(v_1 \wedge \cdots \wedge v_k) \wedge (w_1 \wedge \cdots \wedge w_\ell) = v_1 \wedge \cdots \wedge v_k \wedge w_1 \wedge \cdots \wedge w_\ell.$$

This defines the **exterior algebra** $\Lambda^\cdot V = \bigoplus_{d \geq 0} \Lambda^d V$ as a (skew-commutative) ring, where for $\eta \in \Lambda^k V$ and $\xi \in \Lambda^\ell V$, we have $\eta \wedge \xi = (-1)^{k\ell} \xi \wedge \eta$. (It is known that $\dim \Lambda^\cdot V = 2^{\dim V}$.)

3.9 Volume and Determinant

If $\dim V = n$, then $\dim \Lambda^n V = 1$. If $(e_1, \ldots, e_n)$ is a basis of V, then $e_1 \wedge \cdots \wedge e_n \in \Lambda^n V$. A choice of isomorphism $\Lambda^n V \xrightarrow{\sim} k$ is determined by the data of a volume form $\mathrm{vol} \in \Lambda^n V^* = (\Lambda^n V)^*$, where $\mathrm{vol} \neq 0$, i.e., a non-degenerate alternating multilinear map:

$$V \times \cdots \times V \to k, \quad (v_1, \ldots, v_n) \mapsto \mathrm{vol}(v_1, \ldots, v_n).$$

This can be interpreted as the signed volume of the parallelepiped with edge vectors $v_1, \ldots, v_n$, which is naturally given by $v_1 \wedge \cdots \wedge v_n \in \Lambda^n V$ and becomes a scalar once we identify $\Lambda^n V \xrightarrow{\sim} k$.

Example 3.75. *In a real inner product space with orthonormal basis $(e_1, \dots, e_n)$, the natural volume form is $vol = e_1^* \wedge \cdots \wedge e_n^*$, so $vol(e_1, \dots, e_n) = 1$. Reordering the basis results in ± 1, indicating that orientation matters!*

If $v_j = \begin{pmatrix} v_{1j} \\ \vdots \\ v_{nj} \end{pmatrix}$ *for each j, we have:*

$$vol(v_1, \dots, v_n) = (e_1^* \wedge \cdots \wedge e_n^*)(v_1, \dots, v_n) = \sum_{\sigma \in S_n} (-1)^\sigma (e_{\sigma(1)}^* \otimes \cdots \otimes e_{\sigma(n)}^*)(v_1, \dots, v_n).$$

This simplifies to the determinant of the matrix formed by the columns $v_1, \dots, v_n$:

$$vol(v_1, \dots, v_n) = det(v_1, \dots, v_n).$$

Recall that the determinant of a matrix is given by:

$$\det(A) = \sum_{\sigma \in S_n} (-1)^\sigma \prod a_{\sigma(j)j}.$$

The determinant is characterized by the following properties:

- It is multilinear in the columns of the matrix.
- It is alternating (i.e., swapping two columns changes the sign of the determinant).
- $\det(\mathrm{Id}) = 1$.

Even though the notion of the determinant/volume of $n = \dim V$ vectors requires a choice of volume form (isomorphism $\Lambda^n V \xrightarrow{\sim} k$), the determinant of a linear operator requires no such choice.

Definition 3.76. *Given a linear map $T : V \to V$, define the determinant of T as $\det(T) = \det(A)$, where $A = \mathcal{M}(T)$ is the matrix representation of T in any basis. Using the property $\det(AB) = \det(A)\det(B)$, we have that under a change of basis, $\det\left(P^{-1}AP\right) = \det(A)$.*

Definition 3.77. *The exterior power is a functor. Given a linear map $T : V \to V$, it induces a linear operator $\Lambda^n T : \Lambda^n V \to \Lambda^n V$, defined explicitly by*

$$(\Lambda^n T)(v_1 \wedge \cdots \wedge v_n) = T(v_1) \wedge \cdots \wedge T(v_n).$$

Since $\dim(\Lambda^n V) = 1$, and any linear operator on a 1-dimensional vector space is a scalar multiple of the identity map, we define $\det(T) \in k$ such that

$$\Lambda^n T = \det(T)\, id.$$

This definition expresses the fact that the map T scales the volume of parallelepipeds in V by a factor of $\det(T)$, without the need to choose an isomorphism $\Lambda^n V \cong k$ to measure the volume.

By defining the determinant in terms of the action on the n-th exterior power, the independence of the choice of basis is immediate, and so is the property that $\det(T_1 T_2) = \det(T_1)\det(T_2)$ for any linear maps T_1 and T_2.

4 Group Theory II

4.1 Modules

Let R be a commutative ring (with $1 \neq 0$), i.e., relaxing the field axioms by not requiring multiplicative inverses. Main examples include $R = \mathbb{Z}, \mathbb{Z}/n, k[x], k[x_1, \ldots, x_n]$.

Definition 4.1. *A **module** M over a ring R is a set with two operations:*

- $+ : M \times M \to M$ *(addition), such that $(M, +)$ is an abelian group.*
- $\times : R \times M \to M$ *(scalar multiplication), satisfying:*

$$(ab)v = a(bv), \quad a(v+w) = av+aw, \quad (a+b)v = av+bv, \quad 0v = 0, \quad 1v = v.$$

Example 4.2.

- $R^n = \{(x_1, \ldots, x_n) \mid x_i \in R\}$ *with component-wise operations is the **free module** of rank n over R.*
- *Any abelian group is a $\mathbb{Z}$-module (where $n \cdot g = g + \cdots + g$ n times).*

Definition 4.3.

- *A subset $\Gamma \subset M$ **spans** M (or is a **generating set**) if every element of M is a finite linear combination $\sum a_i v_i$, where $v_i \in \Gamma$ and $a_i \in R$. Equivalently, the map $\varphi : R^\Gamma \to M$, defined by $\varphi((a_i)) = \sum a_i v_i$, is surjective. The module M is **finitely generated** if it has a finite spanning set.*
- *The elements of $\Gamma \subset M$ are **linearly independent** if $\varphi : R^\Gamma \to M$ is injective, i.e., $\sum a_i v_i = 0$ with $v_i \in \Gamma$ and $a_i \in R$ implies $a_i = 0$ for all i.*
- *The elements of $\Gamma \subset M$ form a **basis** if $\varphi : R^\Gamma \to M$ is an isomorphism. In this case, M is called a **free module**.*

In general, nothing is true for modules!

- A basis does not need to exist! For example, consider $M = \mathbb{Z}/n$ as a $\mathbb{Z}$-module: $nx = 0$ for all $x \in M$, so $\varphi : \mathbb{Z}^\Gamma \to M$ cannot be injective.
- Even if M is free (admits a basis), a linearly independent set may not be a subset of a basis. For instance, consider $M = \mathbb{Z}$ as a $\mathbb{Z}$-module. No basis contains 2 as an element. Similarly, a spanning set does not need to contain a subset that is a basis. For example, in $M = \mathbb{Z}$ as a $\mathbb{Z}$-module, $\{4, 5\}$ spans $\mathbb{Z}$ (since $n = n \cdot 5 - n \cdot 4$) but is not independent ($5 \cdot 4 - 4 \cdot 5 = 0$). Neither subset $\{4\}$ nor $\{5\}$ spans all of $\mathbb{Z}$.
- A submodule of a finitely generated module need not be finitely generated. For example, let $R = k[x_1, x_2, \ldots]$ (polynomials in infinitely many variables) and $M = R$ as an R-module generated by 1. The submodule $M' = \{\text{polynomials with zero constant term}\} \subset M$ is not finitely generated because any finite subset involves only finitely many x_i's, which

cannot span the remaining variables. By contrast, this holds for modules over Noetherian rings, including $\mathbb{Z}$ and $k[x_1, \ldots, x_n]$.

Definition 4.4. *Let M, N be modules over R. A **module homomorphism** $\varphi \in Hom_R(M, N)$ is a map $\varphi : M \to N$ such that:*

$$\varphi(v + w) = \varphi(v) + \varphi(w), \quad \varphi(av) = a\varphi(v).$$

Observe that $\mathrm{Hom}_R(M, N)$ is itself an R-module. For free modules, things work as expected: $\mathrm{Hom}_R(R^m, R^n) \simeq R^{m \times n}$ (since φ is determined by the images $\varphi(e_i) \in R^n$ of the basis vectors of R^m). However, nonzero modules M, N can exist such that $\mathrm{Hom}_R(M, N) = 0$.

Example 4.5. *Let $R = k[x]$ and $M = k$, with multiplication defined by $(a_0 + a_1 x + \cdots) \cdot b = a_0 b$. Then $Hom_R(k, k[x]) = 0$ because $1 \in k$ satisfies $x \cdot 1 = 0$, so $\varphi(1) = p(x) \in k[x]$ must satisfy $xp(x) = 0$, implying $p(x) = 0$.*

A couple remarks:

Remark 4.6.

- *R is a module over itself (a free module of rank 1). A submodule of R is called an **ideal**, i.e., a subset $N \subset R$ such that N is an abelian subgroup of $(R, +)$ and $R \times N \subseteq N$ (multiplication by any element of R preserves N). Examples include:*
  - *Ideals in $\mathbb{Z}$: $n\mathbb{Z}$.*
  - *Ideals in $k[x]$: $p(x)k[x]$.*

  Both are generated by a single element, a property special to principal ideal domains (PIDs). This relates to Euclidean division algorithms: $span(p, q) = span(gcd(p, q))$.
- *The quotient of an R-module by a submodule is an R-module. For example:*

  $$\mathbb{Z}/n\mathbb{Z} = \mathbb{Z}/n \text{ as a } \mathbb{Z}\text{-module}, \quad k[x]/xk[x] = k \text{ as a } k[x]\text{-module}.$$

  The quotient of R by an ideal is not only an R-module but also a ring.

The study of modules is a vast subject, which we won't study further, with one exception: we're returning to group theory, but we will start with the classification of finite generated abelian groups (which are $\mathbb{Z}$-modules).

The main theorem is below:

Proposition 4.7. *Any finitely generated abelian group is isomorphic to a product of cyclic groups:*

$$G \simeq (\mathbb{Z}/n_1 \times \cdots \times \mathbb{Z}/n_k) \times \mathbb{Z}^\ell,$$

where, using the fact that $\mathbb{Z}/mn \simeq \mathbb{Z}/m \times \mathbb{Z}/n$ if and only if $gcd(m, n) = 1$, the finite factors can be arranged such that each n_i is a power of a prime.

4.2 Classification of Finitely Generated Abelian Groups

A major challenge with modules is that bases do not always exist, and linearly independent families cannot always be extended to form a basis.

Proposition 4.8. *If M is a finitely generated $\mathbb{Z}$-module, then $\exists m, n$ and $T \in Hom(\mathbb{Z}^m, \mathbb{Z}^n)$ such that $M \simeq \mathbb{Z}^n / Im(T)$. Equivalently, $\exists$ an exact sequence*

$$\mathbb{Z}^m \xrightarrow{T} \mathbb{Z}^n \longrightarrow M \longrightarrow 0.$$

This relies on the following:

Lemma 4.9. *Any submodule of $\mathbb{Z}^n$ is finitely generated (in fact, free of rank $\leq n$).*

Proof. We proceed by induction on n.

For $n = 1$: subgroups of $(\mathbb{Z}, +)$ are either $\{0\}$ or $\mathbb{Z}a$ for $a \in \mathbb{Z} \setminus \{0\}$.

Assume the result holds for $\mathbb{Z}^{n-1}$, and consider a submodule $M \subset \mathbb{Z}^n$. Define the projection map

$$\pi : \mathbb{Z}^n \to \mathbb{Z}^{n-1}, \quad (a_1, \ldots, a_n) \mapsto (a_2, \ldots, a_n).$$

The image $\mathrm{Im}(\pi)$ is a submodule of $\mathbb{Z}^{n-1}$, which is finitely generated (and free) by the induction hypothesis. The kernel $\mathrm{Ker}(\pi) = M \cap (\mathbb{Z} \times 0 \times \cdots \times 0)$ is a subgroup of $\mathbb{Z}$, hence free (of rank 0 or 1).

Since both $\mathrm{Ker}(\pi)$ and $\mathrm{Im}(\pi)$ are finitely generated and free, M must also be finitely generated and free. Let $\{e_1, \ldots, e_k\}$ be a basis for $\mathrm{Ker}(\pi)$, and let $\{f_1, \ldots, f_m\}$ be a generating set for $\mathrm{Im}(\pi)$. Then for any $x \in M$, we can write

$$\pi(x) = \sum a_i f_i \quad \text{for some } a_i \in \mathbb{Z}.$$

Thus, $x - \sum a_i f_i \in \mathrm{Ker}(\pi)$, implying

$$x \in \mathrm{span}(e_1, \ldots, e_k, f_1, \ldots, f_m).$$

Therefore, (e_i, f_j) generate M. □

Now, let's prove the theorem.

Proof. If M is finitely generated with generators $(e_1, \ldots, e_k)$, define $\varphi : \mathbb{Z}^n \to M$ by

$$\varphi(a_1, \ldots, a_n) = \sum a_i e_i.$$

The map φ is surjective, and $\mathrm{Ker}(\varphi) = \mathrm{Im}(T)$ for some $T : \mathbb{Z}^m \to \mathbb{Z}^n$. This gives an exact sequence

$$\mathbb{Z}^m \xrightarrow{T} \mathbb{Z}^n \xrightarrow{\varphi} M \longrightarrow 0,$$

with $\mathrm{Ker}(\varphi) = \mathrm{Im}(T)$. Hence, $M \simeq \mathbb{Z}^n / \mathrm{Im}(T)$. □

The next ingredient is the notion of divisibility for elements of $\mathbb{Z}^n$ (viewed as a free $\mathbb{Z}$-module).

Definition 4.10. *The **divisibility** of a nonzero element $x = (a_1, \ldots, a_n) \in \mathbb{Z}^n$ is the largest $d \in \mathbb{Z}_+$ such that $\exists y \in \mathbb{Z}^n$ with $x = dy$ (i.e., $d = gcd(a_1, \ldots, a_n)$). An element of $\mathbb{Z}^n$ is **primitive** if its divisibility is 1.*

Lemma 4.11. *An element of a free finitely generated $\mathbb{Z}$-module (e.g., $\mathbb{Z}^n$) can be chosen to be part of a basis if and only if it is primitive (or d times a basis element if and only if its divisibility is d).*

Proof. Clearly, elements of a basis $(e_1, \ldots, e_n)$ are primitive, as linear independence prevents $e_i = d\sum a_i e_i$ for any $d > 1$.

For the converse, let $v = a_1 e_1 + \cdots + a_n e_n$ be primitive. Without loss of generality, assume $a_1 \neq 0$ and $|a_1| = \min\{|a_i| : a_i \neq 0\}$. Using the Euclidean algorithm, redefine the basis to iteratively reduce the coefficients of v, ultimately leaving v as a multiple of a basis vector. □

Proposition 4.12. *For any $T \in Hom(\mathbb{Z}^m, \mathbb{Z}^n)$, there exist bases $(e_1, \ldots, e_m)$ of $\mathbb{Z}^m$, $(f_1, \ldots, f_n)$ of $\mathbb{Z}^n$, $r \leq \min(m, n)$ (the rank of T), and positive integers $d_1, \ldots, d_r$ such that*

$$T(e_i) = \begin{cases} d_i f_i & \text{if } 1 \leq i \leq r, \\ 0 & \text{if } i > r. \end{cases}$$

Equivalently, T can be represented as a block matrix:

$$\left(\begin{array}{ccc|c} d_1 & & 0 & \\ & \ddots & & 0 \\ 0 & & d_r & \\ \hline & 0 & & 0 \end{array}\right).$$

Proof. If $T = 0$, the statement is obvious for all m, n. Otherwise, proceed by induction on m.

Base case $m = 1$**:** Let $d = \text{div}(T(1))$. By the lemma, there exists a basis of $\mathbb{Z}^n$ such that $T(1) = df_1$. Assume the result holds for $\mathbb{Z}^{m-1}$.

Now, consider the case when $m > 1$. Let $T : \mathbb{Z}^m \to \mathbb{Z}^n$ (assume $T \neq 0$). Let $d_1 = \min\{\text{div}(T(x)) \mid x \notin \text{Ker}(T)\}$, and let e_1 be such that $\text{div}(T(e_1)) = d_1$. Note that e_1 must be primitive: if it were divisible by some integer d, then $\text{div}(T(\frac{1}{d}e_1)) = \frac{1}{d}\text{div}(T(e_1))$.

Thus, we can write $T(e_1) = d_1 f_1$, where $f_1 \in \mathbb{Z}^n$ is primitive. Using the lemma, extend e_1 to a basis $(e_1, e_2, \ldots, e_m)$ of $\mathbb{Z}^m$, and $(f_1, f_2, \ldots, f_m)$ of $\mathbb{Z}^n$.

Now, the matrix representation of T with respect to these bases is given by

$$\mathcal{M}(T, (e_i), (f_i)) = \left(\begin{array}{c|c} d_1 & * \\ \hline 0 & \mathcal{M}(T') \end{array}\right)$$

where T' is the restriction of T to $\text{span}(e_2, \ldots, e_m) \simeq \mathbb{Z}^{m-1}$, and T' is composed with the projection to $\text{span}(f_2, \ldots, f_n) \simeq \mathbb{Z}^{n-1}$. By the induction hypothesis, we can replace $(e_2, \ldots, e_m)$ and $(f_2, \ldots, f_n)$ with some other basis of their spans, and assume

$$T'(e_j) = \begin{cases} d_j f_j & \text{for } d_j \leq r \\ 0 & \text{otherwise} \end{cases}$$

Then, the matrix representation becomes

$$\mathcal{M}(T, (e_i), (f_i)) = \left(\begin{array}{c|ccc} d_1 & a_1 & \ldots & a_m \\ \hline & d_2 & & 0 \\ 0 & & \ddots & \\ & 0 & & d_n \end{array}\right)$$

i.e., $T(e_j) = d_j f_j + a_j f_1$ for some $a_j \in \mathbb{Z}$ for $j \geq 2$. Write $a_j = q_j d_1 + r_j$, and change basis to $(e_1, e'_2 = e_2 - q_2 e_1, \ldots, e'_m = e_m - q_m e_1)$. Then the matrix representation becomes

$$\mathcal{M}(T, (e_i), (f_i)) = \left(\begin{array}{c|ccc} d_1 & r_1 & \ldots & r_m \\ \hline & d_2 & & 0 \\ 0 & & \ddots & \\ & 0 & & d_n \end{array}\right)$$

with $0 \leq r_2, \ldots, r_m < d_1$. Now, if $r_j \neq 0$, it would imply $\text{div}(T(e_j)) \mid r_j < d_1$, contradicting our choice of d_1. Therefore, $r_j = 0$ for all $j \geq 2$, and we are done. □

Now we prove the theorem:

Proof. Proposition 1.8 implies that any finitely generated $\mathbb{Z}$-module M is isomorphic to $\mathbb{Z}^n/\text{Im}(T)$ for some $T \in \text{Hom}(\mathbb{Z}^m, \mathbb{Z}^n)$. Proposition 1.12 ensures that, after a suitable change of basis on $\mathbb{Z}^n$, $\text{Im}(T)$ is spanned by $d_1 f_1, \ldots, d_r f_r$ for some $d_i > 0$, $r \leq n$. Thus,

$$M \simeq \mathbb{Z}^n/\text{Im}(T) \simeq \mathbb{Z}/d_1 \times \cdots \times \mathbb{Z}/d_r \times \mathbb{Z}^{n-r}.$$

□

4.3 Group Actions

Definition 4.13. *An **action** of a group G on a set S is a homomorphism $\rho : G \to Perm(S)$. Equivalently, we have a map $G \times S \to S$, $(g, s) \mapsto g \cdot s$, such that:*

- *$e \cdot s = s$ for all $s \in S$, and*
- *$(gh) \cdot s = g \cdot (h \cdot s)$ for all $g, h \in G$ and $s \in S$.*

This gives rise to the idea of groups as symmetries of geometric objects. Understanding the sets on which a group G acts (and the manner of action) provides insight into the structure of G.

Definition 4.14. *An action is **faithful** if ρ is injective.*

Otherwise, the group that "really acts" on S is $G/\ker(\rho)$...

Definition 4.15. *The **orbit** of $s \in S$ under G is defined as*

$$O_s = G \cdot s = \{g \cdot s \mid g \in G\} \subseteq S.$$

Observe: $t \in O_s$ if and only if there exists $g \in G$ such that $g \cdot s = t$. Conversely, $s = g^{-1} \cdot t \in O_s$.

Equivalently, $s \sim t$ if and only if there exists $g \in G$ such that $g \cdot s = t$. This defines an equivalence relation:

- Reflexivity: $s \sim s$ since $e \cdot s = s$.
- Symmetry: If $s \sim t$, then there exists $g \in G$ such that $g \cdot s = t$. Thus, $t = g \cdot s$ implies $s = g^{-1} \cdot t$, so $t \sim s$.
- Transitivity: If $s \sim t$ and $t \sim u$, then there exist $g, h \in G$ such that $g \cdot s = t$ and $h \cdot t = u$. Hence, $(hg) \cdot s = h \cdot (g \cdot s) = u$, so $s \sim u$.

The orbits are the equivalence classes of this relation.

Definition 4.16. *An action is **transitive** if there is only one orbit, i.e., for all $s, t \in S$, there exists $g \in G$ such that $g \cdot s = t$.*

Note: Given any G-action on S, by restriction, we obtain a G-action on each orbit. Each of these actions is transitive (by definition). Thus, any group action can be decomposed into a disjoint union of transitive actions.

Definition 4.17. *The **stabilizer** of $s \in S$ is defined as*

$$Stab(s) = \{g \in G \mid g \cdot s = s\}.$$

This is a subgroup of G.

Definition 4.18. *The **fixed points** of $g \in G$ are the subset*

$$S^g := \{s \in S \mid g \cdot s = s\}.$$

If $s' = g \cdot s$, then $\mathrm{Stab}(s') = g\mathrm{Stab}(s)g^{-1}$. This implies the following proposition:

Proposition 4.19. *Elements in the same orbit have conjugate stabilizers.*

Proof. If $h \cdot s = s$, then $(ghg^{-1}) \cdot (g \cdot s) = g \cdot (h \cdot s) = g \cdot s$. Thus, $g\mathrm{Stab}(s)g^{-1} \subseteq \mathrm{Stab}(s')$. Conversely, applying the same argument to $s = g^{-1} \cdot s'$ gives $g^{-1}\mathrm{Stab}(s')g \subseteq \mathrm{Stab}(s)$. Hence, equality holds. □

Example 4.20. *Let $H \subseteq G$ be a subgroup. Define $G/H = \{\text{cosets } aH\}$. To avoid notation confusion, we write $[H], [aH], \ldots$ for elements of G/H. The group G acts on G/H by left multiplication: $g \cdot [aH] = [gaH]$. This action is transitive, since ba^{-1} maps $[aH]$ to $[bH]$. Furthermore:*

- *$Stab([H]) = H$, and*
- *$Stab([aH]) = aHa^{-1}$.*

The following is what a general group action looks like when restricted to an orbit.

Proposition 4.21. *If G acts on a set S, and $s \in S$, let $H = Stab(s) \subseteq G$. Then*

$$\epsilon : G/H \to O_s, \quad [aH] \mapsto a \cdot s$$

is a bijection and equivariant, i.e., it intertwines the G-actions:

$$\epsilon(g \cdot [aH]) = g \cdot \epsilon([aH]).$$

Note that this is

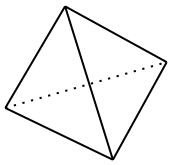
- **Well-defined:** If $a' = ah \in aH$, then $a' \cdot s = a \cdot (h \cdot s) = a \cdot s$.
- **Surjective:** By definition of the orbit.
- **Injective:** $a' \cdot s = a \cdot s$ implies $a^{-1}(a' \cdot s) = s$. Hence $a^{-1}a' \in \mathrm{Stab}(s) = H$, so $a' \in aH$.

For example, the action of G on the orbit O_s is the same as on $G/\mathrm{Stab}(s)$ and the action of G on S is obtained as a disjoint union over orbits.

Corollary 4.22. *If G and S are finite, then*

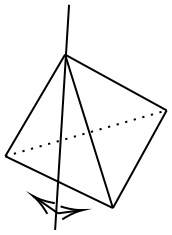
$$|O_s| = \frac{|G|}{|Stab(s)|}, \quad |S| = \sum |O_s|.$$

Example 4.23. *Let G be the group of rotational symmetries of a tetrahedron acting on the set S of its faces, where $|S| = 4$.*



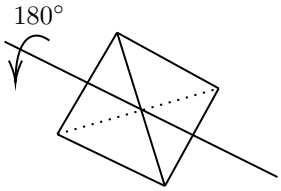
The action is transitive, i.e., there is only one orbit, $|O| = |S| = 4$. The stabilizer of an element $A \in S$ consists of the rotations that map a face to itself, which implies $|Stab(A)| = 3$. Therefore, we find $|G| = |O_A|$ and $|Stab(A)| = 4 \cdot 3 = 12$.

In fact, $G \simeq A_4 \subset S_4$: id and we have 8 elements of order 3:



120°

as well as 3 *elements of order* 2*:*



Now, let's take a look at Burnside's Lemma, a formula to count orbits of a group action. Let G be a finite group acting on a finite set S. Consider $\sum = \{(g, s) \in G \times S \mid g \cdot s = s\}$. There are two ways of calculating $|\sum|$:

1. As a sum over G: $|\sum| = \sum_{g \in G} |S^g|$.
2. As a sum over S: $|\sum| = \sum_{s \in S} |\mathrm{Stab}(s)|$.

However, since all elements in an orbit O have conjugate stabilizers of size $|\mathrm{Stab}(s)| = |G|/|O|$ as seen above, we can rewrite this by grouping over orbits:

$$\begin{aligned} |\sum| &= \sum_{s \in S} |\mathrm{Stab}(s)| \\ &= \sum_{O \text{ orbit}} (|O| \cdot |\mathrm{Stab}(O)|) \\ &= \sum_{O \text{ orbit}} |G| \cdot \frac{|G|}{|O|} \\ &= |G| \cdot (\text{number of orbits}). \end{aligned}$$

This implies Burnside's Lemma.

Proposition 4.24 (Burnside's Lemma)**.** *The number of orbits is equal to* $\frac{1}{|G|} \sum_{g \in G} |S^g|$.

Example 4.25. *How many ways are there to color the faces of a tetrahedron with 3 colors, up to symmetries?*

Let S *be the set of all colorings of the faces, so* $|S| = 3^4 = 81$*. Let* $G = A_4$ *be the group of rotations of the tetrahedron.*

- e *= identity:* $|S^g| = |S| = 81$*.*

- $120°$ *rotation: There are 8 such elements g, and 3 sides have the same color* $\implies |S^g| = 3 \times 3 = 9$.
- $180°$ *rotation: There are 3 such elements g, and* $|S^g| = 3 \times 3 = 9 \implies n = \frac{1}{|G|}\sum_{g\in G}|S^g| = \frac{1}{12}(81 + 11 \cdot 9) = 15$.

Now, let's take a look at Burnside's Lemma, a formula to count orbits of a group action. Let G be a finite group acting on a finite set S. Consider $\sum = \{(g, s) \in G \times S \mid g \cdot s = s\}$. There are two ways of calculating $|\sum|$.

1. As a sum over G: $|\sum| = \sum_{g\in G}|S^g|$.
2. As a sum over S: $|\sum| = \sum_{s\in S}|\mathrm{Stab}(s)|$.

However, since all elements in an orbit O have conjugate stabilizers of size $|\mathrm{Stab}(s)| = |G|/|O|$ as seen above, we can rewrite this by grouping over orbits:

$$|\sum| = \sum_{s\in S}|\mathrm{Stab}(s)| = \sum_{O \text{ orbit}}(|O|\cdot|\mathrm{Stab}(O)|) = \sum_{O \text{ orbit}}|G|\cdot\frac{|G|}{|O|} = |G|\cdot(\text{number of orbits}).$$

This implies Burnside's Lemma.

Proposition 4.26 (Burnside's Lemma)**.** *The number of orbits is equal to* $\frac{1}{|G|}\sum_{g\in G}|S^g|$.

Example 4.27. *How many ways are there to color the faces of a tetrahedron with 3 colors, up to symmetries?*

Let S be the set of all colorings of the faces, so $|S| = 3^4 = 81$. Let $G = A_4$ be the group of rotations of the tetrahedron.

- *e = identity:* $|S^g| = |S| = 81$.
- $120°$ *rotation: There are 8 such elements g, and 3 sides have the same color* $\implies |S^g| = 3 \times 3 = 9$.
- $180°$ *rotation: There are 3 such elements g, and* $|S^g| = 3 \times 3 = 9 \implies n = \frac{1}{|G|}\sum_{g\in G}|S^g| = \frac{1}{12}(81 + 11 \cdot 9) = 15$.

Now, let's look at actions of G on itself.

If G acts on itself by left multiplication: $g \cdot h = gh$. This is transitive, with $\mathrm{Stab}(h) = \{e\}$ for all $h \in G$. The fixed points are φ for all $g \neq e$. The map $G \hookrightarrow \mathrm{Perm}(G)$ is faithful. Thus, we get

Proposition 4.28 (Cayley's Theorem)**.** *Every finite group G is isomorphic to a subgroup of S_n, where $n = |G|$.*

This is not particularly useful for understanding G, however. Here is a more useful action.

If G acts on itself by conjugation: g acts by $h \mapsto ghg^{-1}$.

We've seen that this defines a homomorphism $G \to \mathrm{Aut}(G) \subset \mathrm{Perm}(G)$, so it is indeed an action. Now we have a more interesting structure: the orbits of this action are the conjugacy classes in G, and the stabilizer of an element $h \in G$ is $\mathrm{Stab}(h) = \{g \in G \mid gh = hg\}$ ($ghg^{-1} = h \iff gh = hg$), the subgroup of elements which commute with h. This is called the **centralizer** of h, denoted $Z(h) \subset G$. Note that $\bigcap_{h \in G} Z(h) = Z(G)$, the center of G, is the kernel of the action (i.e., the subgroup of elements which act trivially).

Thus, the action is trivial when G is abelian, and faithful if and only if $Z(G) = \{e\}$. How does this help?

The conjugacy classes form a partition of G, so

$$|G| = \sum_{C \subset G} |C|,$$

which is called the **class equation** of the group G.

For each conjugacy class, $|C_h| = \frac{|G|}{|Z(h)|}$ divides G. Moreover, $|C_e| = 1$ for the identity element, and $|C_h| = 1$ if and only if $h \in Z(G)$.

This is extremely useful. For example:

Theorem 4.29. *If $|G| = p^2$ for some prime p, then G must be abelian.*

Proof.

- Conjugacy classes have orders $|C| \in \{1, p, p^2\}$ and $\sum |C| = p^2$. Thus, the number of conjugacy classes such that $|C| = 1$ (i.e., of central elements of G) must be a multiple of p. Hence, $p \mid |Z(G)|$.
- $Z(G)$ is a subgroup of G, so $|Z(G)|$ divides p^2: it must be either p or p^2. If $|Z(G)| = p^2$, then G is abelian.
- Now assume $|Z(G)| = p$ and let $g \notin Z(G)$. Then g commutes with itself and with $Z(G)$, so $Z(g) \supset Z(G) \cup \{g\}$, hence $|Z(g)| > p$. But $Z(g)$ is a subgroup of G, so $|Z(g)| \mid p^2$. This implies $Z(g) = G$, i.e., g commutes with all elements of G, i.e., $g \in Z(G)$, which is a contradiction. Thus, $Z(G) = G$, and G is abelian.

□

Proposition 4.30. *There are exactly 5 groups of order 8 up to isomorphism.*

These are $\mathbb{Z}/8$, $\mathbb{Z}/2 \times \mathbb{Z}/4$, $(\mathbb{Z}/2)^3$, D_4, and the quaternions.

There are two ways to show that there are exactly two non-abelian groups of order 8.

1. **By hand**: If $|G| = 8$ and G is not abelian, then
   (a) A group where every element has order 2 must be abelian, so there must be an element a of order 4.

(b) The order 4 subgroup generated by a is normal. Work out the possibilities for multiplication by an element b such that $ab \neq ba$.

2. **Using conjugacy and the class equation**:

(a) The class equation is $8 = \sum |C|$, with $|C| \in \{1, 2, 4, 8\}$. Since $|C_e| = 1$, we have $Z(G) = \{g \mid |C_g| = 1\}$, and its order must be 2, 4, or 8. It cannot be 8 because then G would be abelian, and 4 is impossible by the same argument as for p^2 above. Thus, $|Z(G)| = 4$.

(b) If $g \notin Z(G)$, then $Z(g) \subsetneq G$, but $Z(G) \cup \{g\} \subset Z(g)$. Hence, $|Z(g)| = 4$ and $|C_g| = 2$. Therefore, the class equation is $8 = 1+1+2+2+2$. Now, work out the possibilities.

4.4 Finite Subgroups of SO(3)

Let us use group actions to classify the finite subgroups of $SO(3)$, the group of rotations of $\mathbb{R}^3$.

Recall: for an inner product space $(V, \langle ., . \rangle)$, the orthogonal group is defined as $O(V) = \{T \in GL(V) \mid \langle Tu, Tv \rangle = \langle u, v \rangle \ \forall \ u, v \in V\}$. The elements of $O(V)$ satisfy $\det(T) = \pm 1$, and the special orthogonal group is defined as $SO(V) = \{T \in O(V) \mid \det(T) = 1\}$ ("the connected component of Id in $O(V)$").

We have seen that for $T \in O(V)$, there exists a decomposition $V = \bigoplus V_i$, where $V_i \perp V_j$ for $i \neq j$, $\dim(V_i) \in \{1, 2\}$, and $T(V_i) = V_i$. This result follows from the fact that there exists an initial invariant subspace, and if W is invariant, then so is $W^\perp$. If $\dim(V_i) = 1$, then $T_{|V_i} = \pm 1$, and if $\dim(V_i) = 2$, then $T_{|V_i}$ is a rotation.

In dimension 3, either

$$T \sim \begin{bmatrix} \pm 1 & & \\ & \pm 1 & \\ & & \pm 1 \end{bmatrix} \quad \text{or} \quad T \sim \left[\begin{array}{c|c} \pm 1 & \\ \hline & \text{rotation} \end{array}\right].$$

The condition $\det(T) = 1$ restricts the possibilities to Id,

$$\begin{bmatrix} 1 & & \\ & -1 & \\ & & 1 \end{bmatrix}, \quad \text{and} \quad \left[\begin{array}{c|c} 1 & \\ \hline & \text{rotation} \end{array}\right].$$

This implies that every element of $SO(3)$ is a rotation. If $T \neq \text{Id}$, then T has an axis (the $+1$-eigenspace, which is a line), and it rotates by some angle in the plane perpendicular to the axis.

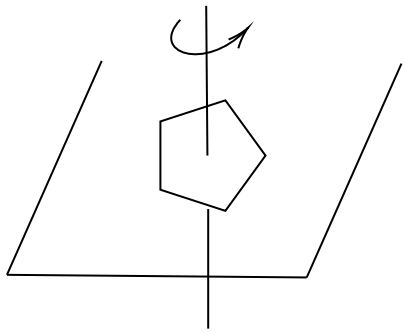
Given a subset $\Sigma \subset \mathbb{R}^3$, we can consider its symmetry group:

$$\{T \in SO(3) \mid T(\Sigma) = \Sigma\}.$$

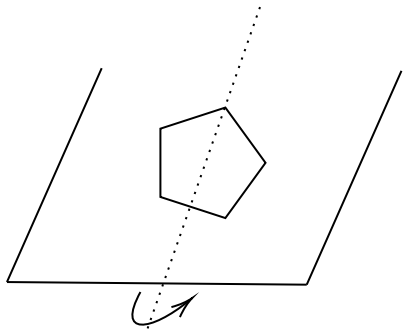
This symmetry group can either be infinite (e.g., if Σ is a circle in a plane, all rotations with an axis perpendicular to the plane will be symmetries) or finite.

Example 4.31. *Let Σ be a regular n-gon in a plane (centered at the origin).*

- *The group contains n rotations (with axis perpendicular to the plane and angles $\frac{2\pi k}{n}$):*



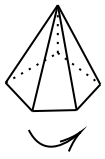
- *The group also contains n flips, which are rotations by π about axes lying in the plane. Together, this group is isomorphic to the dihedral group D_n:*



There exist some special cases:

$n = 1$ $\qquad$ $n = 2$

Example 4.32. *To only keep $\mathbb{Z}/n \subset D_n$ in the above example, consider a cone on a regular n-gon in a plane:*



Example 4.33. *Symmetry of regular polyhedra:*

- *The tetrahedron, cube, and octahedron have the same symmetries, so by*

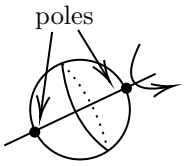
duality, vertices $\Longleftrightarrow$ centers of the faces. In other words, $P^* = \{v \in \mathbb{R}^3 \mid \langle v, u\rangle \leq 1 \ \forall u \in P\}$.

- *The dodecahedron (with 20 pentagonal faces) and icosahedron (with 20 triangular faces) are duals and share the same symmetries.*

These correspond to the symmetry groups A_4 (action on vertices or faces of the tetrahedron), S_4 (action on the 4 diagonals of the cube), and A_5 (action on the icosahedron).

Theorem 4.34. *This is the complete list of finite subgroups of $SO(3)$: $\mathbb{Z}/n, D_n, A_4, S_4, A_5$.*

The key observation is that every $T \in SO(3)$, $T \neq \mathrm{id}$, is a rotation about some axis, hence fixes exactly two unit vectors, $\pm v$, called the **poles** of T:



For a finite subgroup $G \subset SO(3)$, let P denote the set of all poles of elements of G:

$$P = \{v \in \mathbb{R}^3 \mid \|v\| = 1 \text{ and } \exists g \in G, g \neq \mathrm{id} \text{ such that } gv = v\}.$$

Now, if v is a pole of $g \in G$, and given any $h \in G$, $h(v)$ is a pole of $hgh^{-1} \in G$ (since $hgh^{-1} \cdot hv = hgv = hv$). Hence, G acts on P! This is the key to understanding the group G.

Example 4.35. *In the case of symmetry groups of regular polyhedra, we have $P = \{\text{vertices}\} \cup \{\text{centers of faces}\} \cup \{\text{midpoints of edges}\}$. These form three distinct orbits under the action of G on P: one orbit for the vertices, one for the faces, and one for the edges. (In the case of a regular polyhedron, each of these actions is transitive.)*

The next observation is that for $p \in P$, the stabilizer $\mathrm{Stab}(p)$ consists of rotations with axis $\pm p$. These form an abelian cyclic subgroup of G. Thus, $\mathrm{Stab}(p) \simeq \mathbb{Z}/r_p$, where $r_p > 1$ (since p is a pole of some element of G, $\mathrm{Stab}(p)$ must be nontrivial). In other words, the angles of rotation about p form a finite subgroup of $\mathbb{R}/2\pi\mathbb{Z}$, and these angles must be multiples of $\frac{2\pi}{r_p}$. With this understanding, the proof of the theorem follows as a counting argument.

Proof. et $G \subset SO(3)$ be a nontrivial finite subgroup, and let P be the set of poles as defined above. Define $\sum = \{(g, p) \in G \times P \mid g \neq e, g(p) = p\}$ (i.e., p is a pole of g). For each element of $G \setminus \{e\}$, there are exactly 2 poles. Therefore, $|\sum| = 2|G| - 2$. For each $p \in P$, there are $r_p - 1$ rotations in $G \setminus \{e\}$ that fix p. Thus,

$$|\sum| = 2|G| - 2 = \sum_{p \in P} (r_p - 1).$$

Now, the elements $p \in O$ of an orbit of G have conjugate stabilizers ($\text{Stab}(g(p)) = g\text{Stab}(p)g^{-1}$), and hence the stabilizers $\text{Stab}(p)$ have the same order: $r_{gp} = r_p$. Thus, we have:

$$2|G| - 2 = \sum_{O_i \text{ orbit}} |O_i|(r_i - 1),$$

where $r_i = r_p$ for $p \in O_i$.

Recall the orbit-stabilizer theorem: $|O_i| = \frac{|G|}{\text{Stab}} = \frac{|G|}{r_i}$, so we obtain:

$$2|G| - 2 = \sum_{O_i \text{ orbit}} \frac{|G|}{r_i}(r_i - 1),$$

which simplifies to:

$$2 - \frac{2}{|G|} = \sum_{\text{orbits}} 1 - \frac{1}{r_i}.$$

The right-hand side increases rapidly if there are too many orbits: each term is at least $\frac{1}{2}$ since $r_i \geq 2$. Therefore, the number of orbits is at most 3.

Now, we analyze each case based on the number of orbits:

- **1 orbit**: This is impossible because the left-hand side is at least 1 (since $|G| \geq 2$), while the right-hand side is less than 1.
- **2 orbits**: We have:

  $$2 - \frac{2}{|G|} = 1 - \frac{1}{r_1} + 1 - \frac{1}{r_2},$$

  which simplifies to:

  $$\frac{2}{|G|} = \frac{1}{r_1} + \frac{1}{r_2}.$$

  Since each $r_i = |\text{Stab}(p)|$ divides $|G|$, we must have $r_1 = r_2 = |G|$. Hence, $\text{Stab} = G$, and there are exactly two poles, $\pm p$, each fixed under all of G. Therefore, $G = \text{Stab}(p)$, and G is a cyclic subgroup consisting of rotations by $\frac{2\pi k}{r}$ about the axis $\pm p$.
- **3 orbits**: We have:

  $$2 - \frac{2}{|G|} = 3 - \frac{1}{r_1} - \frac{1}{r_2} - \frac{1}{r_3}.$$

  Assume $2 \leq r_1 \leq r_2 \leq r_3$. Then:

  $$\frac{1}{r_1} + \frac{1}{r_2} + \frac{1}{r_3} = 1 + \frac{2}{|G|} > 1,$$

  which implies that $r_1 = 2$ and $r_2 \leq 3$ (otherwise, the sum would be less than or equal to 1). We now analyze two cases for r_2.

1. If $r_2 = 2$, then:

$$2 - \frac{2}{|G|} = 3 - \frac{1}{2} - \frac{1}{2} - \frac{1}{r_3} \implies r_3 = \frac{|G|}{2}.$$

Thus, $|O_3| = \frac{|G|}{r_3} = 2$, so these two poles form an orbit. These poles are necessarily $\pm p$, and half of G will consist of rotations by $\frac{2\pi k}{r_3}$ about $\pm p$ (the stabilizer of $\pm p$), while the other half of G will consist of rotations by 180° that swap $p \longleftrightarrow -p$. Hence, G is a dihedral group.

2. If $r_2 = 3$, then:

$$\sum \frac{1}{r_i} > 1 \implies r_3 \in \{3, 4, 5\}.$$

These three cases correspond to the tetrahedron, cube, and icosahedron. In each case, we have $\frac{2}{|G|} = \sum \frac{1}{r_i} - 1$, implying that $|G| = 12, 24, 60$.

□

Note that for regular polyhedra:

- Poles at the midpoints of edges have $r = 2$.
- Poles at vertices: $r =$ number of faces meeting at that vertex.
- Poles at the centers of faces: r is the number of edges of the face.

Thus, $\frac{2}{|G|} = \sum \frac{1}{r_i} - 1$ implies $|G| = 12, 24, 60$.

We may wonder: what is the 5-element set that the symmetries of the dodecahedron act on?

The answer is that the 20 vertices of a dodecahedron can be partitioned into 5 sets of 4 vertices, each forming a regular tetrahedron. These 5 tetrahedra can be arranged in two distinct ways, which are mirror images of each other but not related by a rotation. A rotation of the dodecahedron then permutes the 5 tetrahedra:

- Rotations about the centers of faces correspond to 5-cycles (24 of them).
- Rotations about the vertices correspond to 3-cycles (123), etc. (20 of them).
- Half rotations about the midpoints of edges correspond to (12)(34), etc. (15 of them).

4.5 Conjugacy Classes in the Symmetric Group S_n

Definition 4.36. *A k-cycle $\sigma = (a_1 a_2 \dots a_k) \in S_n$ is a permutation that maps $a_1 \mapsto a_2, a_2 \mapsto a_3, \dots, a_k \mapsto a_1$, while leaving all other elements fixed.*

Definition 4.37. *Two cycles are **disjoint** if the subsets of elements they permute are disjoint.*

Remark 4.38. *Disjoint cycles commute.*

Proposition 4.39. *Any permutation in S_n can be expressed as a product of disjoint cycles. This decomposition is unique up to reordering the factors, as disjoint cycles commute.*

Algorithm: To find the disjoint cycle decomposition of a permutation σ, proceed as follows:

- Start with the successive images of 1 under σ. This gives a subset of elements cyclically permuted by σ.
- Consider similar subsets for elements not already included, and repeat the process.

In other words, the disjoint cycles correspond to the restrictions of σ to the **orbits** of $\langle \sigma \rangle \subset S_n$ acting on $\{1, \dots, n\}$.

Proposition 4.40. *Let $\sigma = (a_1 \dots a_k)$ be a k-cycle and $\tau \in S_n$ any permutation. Then, $\tau\sigma\tau^{-1} = (\tau(a_1) \dots \tau(a_k))$.*

Proof. Consider the action of $\tau\sigma\tau^{-1}$ on $\{\tau(a_1), \dots, \tau(a_k)\}$. By calculation:

$$\tau(a_i) \mapsto a_i \mapsto a_{i+1} \mapsto \tau(a_{i+1}),$$

showing that the elements $\{\tau(a_1), \dots, \tau(a_k)\}$ are permuted as claimed. For elements not in $\{a_1, \dots, a_k\}$, $\tau(b) \mapsto b \mapsto b \mapsto \tau(b)$, so they remain fixed. □

Corollary 4.41. *All k-cycles are conjugate in S_n. More generally, two permutations $\sigma, \tau \in S_n$ are conjugate if and only if they have the same cycle lengths in their disjoint cycle decompositions. Hence, conjugacy classes in S_n correspond to the partitions of n.*

This means that the conjugacy classes correspond to the ways of writing n as a sum of positive integers, up to reordering the terms.

Example 4.42. *For $n = 3$, the partitions and the sizes of the conjugacy classes are:*

1. *Identity $(3 = 1 + 1 + 1)$: size 1.*
2. *Transpositions $(3 = 2 + 1)$: size 3.*
3. *3-cycles $(3 = 3)$: size 2.*

Example 4.43. *For $n = 4$, the partitions and the sizes of the conjugacy classes are:*

1. *Identity* $1 + 1 + 1 + 1$*: size* 1.
2. *Transpositions* $2 + 1 + 1$*: size* 6.
3. *Two transpositions* $2 + 2$*: size* 3.
4. *3-cycles* $3 + 1$*: size* 8.
5. *4-cycles* 4*: size* 6.

The class equation of S_4 is:

$$24 = 1 + 3 + 6 + 6 + 8.$$

This helps us identify the normal subgroups of S_4. A subgroup $H \subset S_4$ is normal if and only if $aHa^{-1} = H$ for all $a \in S_4$. Thus, a normal subgroup must be a union of conjugacy classes, must include the identity, and its order must divide $|S_4| = 24$. The possible normal subgroups are:

- $1 + 3 = 4$, corresponding to $\{\text{id}\} \cup \{(ij)(kl)\}$. This is indeed a normal subgroup, isomorphic to $\mathbb{Z}/2 \times \mathbb{Z}/2$.
- $1 + 3 + 8 = 12$, corresponding to $\{\text{id}\} \cup \{(ij)(kl)\} \cup \{\text{3-cycles}\}$. This is the alternating group $A_4 \subset S_4$.

Example 4.44. *For $n = 5$, the partitions and the sizes of the conjugacy classes are:*

1. *Identity* $1 + 1 + 1 + 1 + 1$*: size* 1.
2. *Transpositions* $2 + 1 + 1 + 1$*: size* 10.
3. *Two transpositions* $2 + 2 + 1$*: size* 15.
4. *3-cycles* $3 + 1 + 1$*: size* 20.
5. *3-cycle + transposition* $3 + 2$*: size* 20.
6. *4-cycles* $4 + 1$*: size* 30.
7. *5-cycles* 5*: size* 24.

The class equation of S_5 is:

$$120 = 1 + 10 + 15 + 20 + 20 + 30 + 24.$$

To find normal subgroups of S_5 (besides $\{\text{id}\}$ and S_5), we examine the possible unions of conjugacy classes:

- $1 + 15 + 24 = 40$, corresponding to $\{\text{id}\} \cup \{(ij)(kl)\} \cup \{\text{5-cycles}\}$. However, this is not a subgroup since $(12345)(12)(34) = (135)$.

- $1 + 15 + 20 + 24 = 60$, corresponding to $\{\text{id}\} \cup \{(ij)(kl)\} \cup \{\text{5-cycles}\} \cup \{\text{3-cycles}\}$. This is the alternating group $A_5 \subset S_5$.

4.6 The Alternating Group

Recall that we defined the sign homomorphism $\text{sgn} : S_n \to \{\pm 1\}$ by

$$\text{sgn}\left(\prod_{i=1}^{k} \text{transpositions}\right) = (-1)^k,$$

using the fact that transpositions generate S_n. It remains to verify that this definition is independent of the choice of transpositions. Additionally, we have $\text{sgn}(\sigma) = (-1)^{\text{inversions}}$, where the inversions are defined as

$$\text{inversions} = \{(i, j) \mid 1 \leq i < j \leq n \text{ and } \sigma(i) > \sigma(j)\}.$$

This definition requires verification that sgn is indeed a homomorphism.

Alternatively, consider the following approach: take a vector space $V \simeq \mathbb{R}^n$ with basis $(e_1, \ldots, e_n)$, and associate to each $\sigma \in S_n$ an element of $GL(V)$, specifically the linear map $T_\sigma : V \to V$ defined by $T_\sigma(e_i) = e_{\sigma(i)}$. This construction defines an injective homomorphism $S_n \hookrightarrow GL(n)$, with the image being the subgroup of permutation matrices. The map T_σ has finite order (since σ does), so $\det(T_\sigma) \in \mathbb{R}$ is a root of unity and hence $\det(T_\sigma) \in \{\pm 1\}$. We define $\text{sgn}(\sigma) = \det(T_\sigma)$, which is clearly well-defined and a homomorphism.

Concretely, to compute the sign: the action of $\bigwedge^n T_\sigma$ on $\bigwedge^n V$ maps

$$e_1 \wedge \cdots \wedge e_n \mapsto e_{\sigma(1)} \wedge \cdots \wedge e_{\sigma(n)}.$$

The sign corresponds to the number of transpositions required to return the result to the original order, which agrees with the earlier definition.

Observe that a k-cycle has sign $(-1)^{k-1}$ because $(i_1 \ldots i_k) = (i_1 i_2)(i_2 i_3) \ldots (i_{k-1} i_k)$. Therefore, if $\sigma \in S_n$ has cycle lengths $k_1, \ldots, k_l$ (including 1-cycles), corresponding to the partition $n = k_1 + \cdots + k_l$, then

$$\text{sgn}(\sigma) = (-1)^{\sum(k_i - 1)} = (-1)^{n-l}.$$

Definition 4.45. *The **alternating group** is defined as $A_n = \ker(sgn) \subset S_n$, a normal subgroup of index 2 in S_n.*

Proposition 4.46. *If $C \subset S_n$ is a conjugacy class, then one of the following holds:*

1. *C is odd, so $C \cap A_n = \emptyset$.*
2. *$C \subset A_n$ is a conjugacy class in A_n.*
3. *$C \subset A_n$ splits into two conjugacy classes in A_n.*

In the last two cases, consider $\sigma \in C$ and its centralizer $Z(\sigma) = \{\tau \in S_n \mid \tau\sigma\tau^{-1} = \sigma\}$. If $Z(\sigma) \subset A_n$, then the conjugates of σ by odd permutations are distinct from those by even permutations, resulting in two conjugacy classes in A_n. Otherwise, all conjugates of σ in S_n are conjugates by elements of A_n.

Example 4.47. *Consider $n = 5$:*

$$A_5 = \{id\} \cup \{(ij)(kl)\} \cup \{\textit{3-cycles}\} \cup \{\textit{5-cycles}\}.$$

- *The 3-cycles form a single conjugacy class in A_5 because $(45) \in Z((123))$. Similarly, for $(ij)(kl)$, since $(ij) \in Z((ij)(kl))$.*
- *The 5-cycles split into two conjugacy classes in A_5.*

Thus, the class equation of A_5 is

$$60 = 1 + 15 + 20 + 12 + 12.$$

We now consider the normal subgroups of A_5. Since it is impossible to partition 60 (the order of A_5) in a nontrivial way using union of conjugacy classes including {id}, we deduce:

Proposition 4.48. *A_5 is simple.*

Let $p(n)$ denote the number of partitions of n, i.e.,

$$p(n) = \#\{(a_1, \ldots, a_k) \mid a_1 \geq \cdots \geq a_k, \sum a_i = n\}.$$

Alternatively, let $m_j = \#\{1 \mid a_i = j\}$ (the number of times j appears in the partition). Then

$$p(n) = \#\{(m_1, \ldots, m_n) \in \mathbb{N}^n \mid \sum jm_j = n\}.$$

There is no closed formula for $p(n)$, but it grows faster than any polynomial.

Theorem 4.49 (Hardy-Ramanujan 1918)**.**

$$p(n) \sim \frac{1}{4n\sqrt{3}} \exp\left(\phi\sqrt{\frac{2n}{3}}e\right)$$

Remark 4.50. *This is very hard to prove.*

However, there are recursive formulas and a nice expression for the generating series:

$$f(t) = \sum_{n=0}^{\infty} p(n)t^n = \prod_{j=1}^{\infty} \frac{1}{1-t^j},$$

where the coefficient of t^n in this product is $p(n)$. This is because:

$$\frac{1}{1-t^j} = 1 + t^j + t^{2j} + \ldots,$$

so the coefficient of t^n in the product is the number of ways to write n as the sum of multiples of j for $j = 1, 2, \ldots$, i.e., $n = m_1 + 2m_2 + 3m_3 + \ldots$.

Problem 4.51. *What is the size of the conjugacy class in S_n corresponding to a given partition $n = \sum jm_j$? (That is, m_1 fixed elements, m_2 2-cycles, m_3 3-cycles, etc.)*

Answer: We first need to partition $\{1, \ldots, n\}$ into m_1 subsets of size 1, m_2 subsets of size 2, etc.

There are:

$$\frac{n!}{(1!)^{m_1}(2!)^{m_2}\ldots}$$

ways to do this, because S_n acts transitively on the set of such decompositions, with a stabilizer subgroup $\prod_j (S_j)^{m_j}$, which is the product of permutations that permute only within each subset. However, we don't care about the ordering of the various subsets of a given size. Therefore, we divide by $m_j!$ for each j (since we can permute the m_j subsets of size j). Thus, we obtain:

$$\frac{n!}{\prod_{j\geq 1}(j!)^{m_j} m_j!}.$$

Now, in S_j, there are $(j-1)!$ j-cycles, so the total number of ways to choose the cycles acting on each subset is $\prod((j-1)!)^{m_j}$. Hence, the size of the conjugacy class is:

$$|C| = \frac{n!}{\prod_{j\geq 1}(j^{m_j} \cdot m_j!)}.$$

Remark 4.52. *Can you check by direct calculation that these add up to $n! = |S_n|$?*

Let's now return to the alternating group $A_n = \mathrm{Ker}(\mathrm{sgn} : S_n \to \{\pm 1\})$.

Observe: a k-cycle has sign $(-1)^{k-1}$ (since $(i_1 ... i_k) = (i_1 i_2)(i_2 i_3)...(i_{k-1} i_k)$). So $\sigma \in A_n$ if and only if its cycle decomposition has an even number of cycles of even length.

Next, let's return to the alternating group $A_n = \mathrm{Ker}(\mathrm{sgn} : S_n \to \{\pm 1\})$.

Observe that a k-cycle has sign $(-1)^{k-1}$ (since $(i_1 \ldots i_k) = (i_1 i_2)(i_2 i_3) \ldots (i_{k-1} i_k)$). Thus, $\sigma \in A_n$ if and only if its cycle decomposition has an even number of cycles of even length.

Proposition 4.53. *If $C \subset S_n$ is a conjugacy class, then either $C \cap A_n = \emptyset$ or $C \subset A_n$. In the latter case, either C is a conjugacy class in A_n, or it splits into two conjugacy classes in A_n. Specifically, C is a single conjugacy class in A_n if and only if, given $\sigma \in C$, there exists an odd permutation τ that commutes with σ.*

Proof.

- All elements of C have the same cycle lengths, which implies they have the same sign. Therefore, $C \subset A_n$ or $C \cap A_n = \emptyset$ (since A_n is a normal subgroup of S_n, it is a union of conjugacy classes).

- Assume $C_\sigma = \{g\sigma g^{-1} \mid g \in S_n\} \subset A_n$. Then, split S_n into the two right cosets of A_n: $S_n = A_n \cup A_n \cdot \tau$, where $\operatorname{sgn}(\tau) = -1$. Thus,

Thus,

$$C_\sigma = \{h\sigma h^{-1} \mid h \in A_n\} \cup \{h\tau\sigma\tau^{-1}h^{-1} \mid h\tau \in A_n\tau\}.$$

These two conjugacy classes are either equal or disjoint. They are equal if and only if σ is in the latter conjugacy class, i.e., there exists $g = h\tau$ (an odd permutation) such that $g\sigma g^{-1} = \sigma$, or equivalently, $g\sigma = \sigma g$. □

In other words: for $\sigma \in C$, the centralizer $Z(\sigma) = \{\tau \in S_n \mid \tau\sigma\tau^{-1} = \sigma\}$. If $Z(\sigma) \subset A_n$, then conjugates of σ by odd permutations are different from conjugates by even permutations, forming two conjugacy classes in A_n. If $Z(\sigma) \not\subset A_n$, then all conjugates of $\sigma \in S_n$ are conjugates by elements of A_n.

Example 4.54.

$$A_5 = \{id\} \cup \{(ij)(kl)\} \cup \{\textit{3-cycles}\} \cup \{\textit{5-cycles}\}.$$

The 3-cycles still form a single conjugacy class in A_5, as do the $(ij)(kl)$'s, but the 5-cycles split into two conjugacy classes in A_5.

Thus, the class equation of A_5 is $60 = 1 + 15 + 20 + 12 + 12$.

More generally:

Proposition 4.55. *For $\sigma \in A_n$, the conjugacy class $C = \{g\sigma g^{-1} \mid g \in S_n\}$ splits into two conjugacy classes in A_n if and only if the cycle lengths of σ are all odd and distinct.*

Proof.
- σ commutes with the cycles in its own cycle decomposition. Therefore, every even-length cycle in σ gives an odd permutation in $Z(\sigma)$, implying that C_σ does not split.
- If two odd cycles $(a_1 \dots a_k)$ and $(b_1 \dots b_k)$ of the same length appear in the cycle decomposition of σ, then $(a_1b_1)(a_2b_2)\dots(a_kb_k) \in Z(\sigma)$ is odd. This includes the case where $k = 1$, where we cannot have two fixed points.
- If the cycle lengths are all distinct, then an element of $Z(\sigma)$ must permute each of the corresponding subsets of $\{1, \dots, n\}$. On a j-element subset, $Z((12\dots j))$ is the cyclic subgroup of S_j generated by $(12\dots j)$, which is a subgroup of A_j. Hence, $Z(\sigma) \subset A_n$.

□

Now, let's look for the normal subgroups of A_5. We cannot find a divisor of 60 in any nontrivial way as the union of conjugacy classes, except by taking all of A_5. Therefore:

Proposition 4.56. *A_5 is simple, i.e., its only normal subgroups are $\{id\}$ and itself.*

Theorem 4.57. *A_n is simple for all $n \geq 5$.*

We have already shown A_5 is simple; A_6 follows by a similar argument using the class equation. However, the result is false for A_4 (since $\{\mathrm{id}\} \cup \{(ij)(kl)\} \subset A_4$ is normal). The general case relies on the following lemma:

Lemma 4.58. *A_n is generated by 3-cycles.*

Proof. Induction on n: this is true for $A_3 = \{\mathrm{id}, 3\text{-cycles}\} \subset S_3$. Now assume A_{n-1} is generated by 3-cycles. Let $\sigma \in A_n$. If $\sigma(n) = n$, then σ belongs to a subgroup $\{\tau \in A_n \mid \tau(n) = n\} \simeq A_{n-1}$, so it is a product of 3-cycles by the induction hypothesis. Otherwise, let $i = \sigma(n)$ and j be any element distinct from i and n. Then, $\tau \simeq (jin)\sigma \in A_n$, and $\tau(n) = n$, so by the induction hypothesis, τ is a product of 3-cycles, and therefore so is $\sigma = (ijn)\tau$. □

Furthermore, for $n \geq 5$, 3-cycles form a single conjugacy class in A_n; since $(j_1j_2j_3)$ and $(k_1k_2k_3)$ are conjugates by any permutation $j_1 \mapsto k_1$, and some of these elements lie in A_n. Thus, to prove that a normal subgroup $H \subset A_n$ with $H \neq \{e\}$ is all of A_n, it suffices to show that it contains a 3-cycle.

Now let's prove the theorem.

Proof. Let $H \subset A_n$ be a normal subgroup with $H \neq \{e\}$. As noted earlier, it suffices to show that H contains a 3-cycle (and thus, by conjugation, contains all 3-cycles, implying $H = A_n$). Let $\sigma \in H$ with $\sigma \neq e$. We may assume that σ has prime order by replacing it with some power of τ. Let $m = \mathrm{order}(\sigma)$ and p be a prime divisor of m. Then $\sigma^{\frac{m}{p}} \in H$ has order p. Since the order of σ is the least common multiple (LCM) of its cycle lengths, it follows that σ is a product of disjoint p-cycles. We now analyze the cases depending on the value of p.

- **Case 1: $p \geq 5$.** If $p \geq 5$, we write $\sigma = (i_1 \dots i_p)\tau$, where τ permutes the remaining elements, fixing $i_1, \dots, i_p$. Let $g = (i_4i_3i_2)\tau$, and since H is normal, we have $g\sigma g^{-1} \in H$. Now, consider the commutator $g\sigma g^{-1}\sigma^{-1}$. Using the fact that $\sigma = (i_1 \dots i_p)\tau$, we compute:

  $$g\sigma g^{-1}\sigma^{-1} = (i_4i_3i_2) \circ [(i_1 \dots i_p)\tau] \circ (i_2i_3i_4) \circ (\tau^{-1}(i_p \dots i_1)).$$

  After simplifying the action on the elements $i_1, i_2, \dots, i_5$, we get:

  $$i_1 \mapsto i_1, \quad i_2 \mapsto i_4, \quad i_3 \mapsto i_3, \quad i_4 \mapsto i_5, \quad i_5 \mapsto i_2.$$

  This is a 3-cycle, implying that H contains a 3-cycle.

- **Case 2: $p = 3$.** If σ is a 3-cycle, we are done. Otherwise, assume σ is a product of at least two disjoint 3-cycles. Write $\sigma = (i_1i_2i_3)(i_4i_5i_6)\tau$. Let $g = (i_4i_3i_2)$, and compute the commutator $g\sigma g^{-1}\sigma^{-1}$:

  $$g\sigma g^{-1}\sigma^{-1} = (i_1i_5i_2i_4i_3),$$

  which is a 5-cycle. Since 5-cycles belong to A_n, this reduces to the previous case, where $p \geq 5$, and we conclude that H contains a 3-cycle.

- **Case 3: $p = 2$ and σ is a product of two transpositions.** If $\sigma = (i_1 i_2)(i_3 i_4)$, then σ is not in A_n, as it is an odd permutation. Now let $i_5 \notin \{i_1, \ldots, i_4\}$, and define $g = (i_5 i_3 i_1)$. We compute the commutator $g\sigma g^{-1}\sigma^{-1}$:
$$g\sigma g^{-1}\sigma^{-1} = (i_1 i_5 i_2 i_4 i_3),$$
which is a 3-cycle. This reduces to the first case, where $p \geq 5$, and we conclude that H contains a 3-cycle.

- **Case 4: $p = 2$ and σ is a product of at least three transpositions.** If $\sigma = (i_1 i_2)(i_3 i_4)(i_5 i_6)\tau$, then we again let $g = (i_5 i_3 i_1)$ and compute the commutator:
$$g\sigma g^{-1}\sigma^{-1} = (i_1 i_5 i_3)(i_2 i_4 i_6),$$
which is a product of two 3-cycles. Since products of 3-cycles are in A_n, this reduces to Case 2, where $p = 3$, and we conclude that H contains a 3-cycle.

Thus, in all cases, we have shown that H contains a 3-cycle. By conjugation, this implies that H contains all 3-cycles, and therefore $H = A_n$. □

4.7 The Sylow Theorems

Our next topic, still closely related to understanding finite groups, is the Sylow theorems. If $|G| = n$ and $k \mid n$, there is no general reason for G to contain an element of order k, or even a subgroup of order k: the converse to Lagrange's theorem fails.

Example 4.59.
- *A_4 (resp. A_5) has no subgroups of order 6 (resp. 30).*
- *Such a subgroup, if it existed, would necessarily be normal.*

Now, fix a prime p that divides $|G|$, and write $|G| = p^e m$, where $p \nmid m$.

Definition 4.60. *A subgroup $H \subset G$ of order $|H| = p^e$ is called a **Sylow p-subgroup** of G.*

Theorem 4.61 (Sylow, 1872)**.**
- *For every prime p, a Sylow p-subgroup of G exists.*
- *All Sylow p-subgroups are conjugates of each other: if $H, H' \subset G$ are p-Sylow subgroups, then $\exists g \in G$ such that $H' = gHg^{-1}$. Moreover, any subgroup $K \subset G$ with $|K|$ a power of p is contained in a Sylow p-subgroup.*
- *Let s_p be the number of Sylow p-subgroups of G. Then:*
$$s_p \equiv 1 \pmod{p} \quad \text{and} \quad s_p \mid |G| \quad \text{or equivalently, } s_p \mid m = \frac{|G|}{p^e}.$$

Example 4.62 (Classifying Groups of Order 15)**.** *If $|G| = 15$, then there exist Sylow subgroups $H, K \subset G$ with $|H| = 3$ and $|K| = 5$. The number of such Sylow subgroups is determined as follows:*

$$s_3 \mid 5 \quad \textit{and} \quad s_3 \equiv 1 \pmod 3 \implies s_3 = 1,$$
$$s_5 \mid 3 \quad \textit{and} \quad s_5 \equiv 1 \pmod 5 \implies s_5 = 1.$$

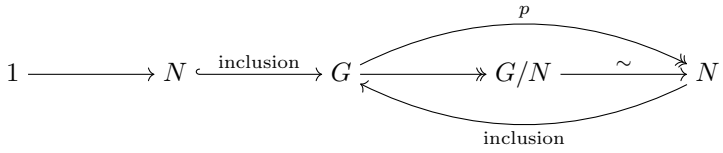
This implies that both H and K are normal (since the conjugates gHg^{-1} and gKg^{-1} are also Sylow subgroups, and H and K are the unique such subgroups). Using the criterion, which we will discuss next, for direct products, this implies:

$$G \simeq H \times K \simeq \mathbb{Z}/3 \times \mathbb{Z}/5 \simeq \mathbb{Z}/15.$$

Thus, every group of order 15 is cyclic.

4.8 (Semi)Direct Products

Let $N \subset G$ be a normal subgroup. Then we have the exact sequence:



where $H \simeq G/N$.

where $H \simeq G/N$.

However, this does not imply that $G \simeq H \times N$, or even that G contains a subgroup isomorphic to H.

Example 4.63. *Consider $\mathbb{Z} \cdot p \subset \mathbb{Z}$ (subgroup), with $0 \to \mathbb{Z}p \to \mathbb{Z} \to \mathbb{Z}/p \to 0$. Here, $\mathbb{Z}$ has no subgroup isomorphic to $\mathbb{Z}/p$.*

On the other hand, assume H can indeed be identified with a subgroup of G via an injective homomorphism $i : H \hookrightarrow G$ such that $p \circ i = \mathrm{id}_H$.

This means N and H are subgroups of G, N is normal, and every coset of N contains a unique element of H. So $H \simeq G/N$, where $h \mapsto hN = Nh$ is a group isomorphism. The above setup arises as:

$$1 \to N \overset{\text{inclusion}}{\to} G \to G/N \simeq H.$$

Thus, every element of G can be uniquely expressed as $g = nh$ with $n \in N$ and $h \in H$. There is therefore a bijection of sets $N \times H \to G$, $(n, h) \mapsto n \cdot h$. This need not be a group isomorphism (particularly because H need not be a normal subgroup of G). However, since N is normal, we have:

$$(n_1 h_1)(n_2 h_2) = (n_1 h_1 n_2 h_1^{-1})(h_1 h_2).$$

This can be interpreted as a semi-direct product of N and H:

Definition 4.64. *Given groups N and H, and an action of H on N by automorphisms (i.e., a homomorphism $\phi : H \to Aut(N)$), we define the* ***semidirect product*** *$N \rtimes_\phi H$ as follows:*

- *Set: $N \times H$.*
- *Group Law $(n_1, h_1) \cdot (n_2, h_2) = (n_1\phi(h_1)(n_2), h_1h_2)$.*

Remark 4.65. *Check that this satisfies the group axioms, particularly associativity.*

In the above setting, $H \subset G$ acts on the normal subgroup $N \subset G$ by conjugation: $\phi(h)(n) = hnh^{-1}$. Then, $G \simeq N \rtimes_\phi H$. To summarize:

Proposition 4.66. *If N and H are subgroups of G, with N normal, such that every coset of N contains a unique element of H (i.e., every element of G is uniquely expressible as $g = n \cdot h$), then G is isomorphic to the semidirect product $N \rtimes_\phi H$.*

Example 4.67.

$$1 \to A_3 \to S_3 \xrightarrow{sgn} \mathbb{Z}/2 \to 1,$$

where $A_3 = \{1, \sigma, \sigma^2\} \simeq \mathbb{Z}/3$ (alternating subgroup, normal). We can realize $\mathbb{Z}/2$ as the subgroup $\{id, \tau\} \subset S_3$ (τ is a transposition, not normal), so:

$$S_3 \simeq \mathbb{Z}/3 \rtimes \mathbb{Z}/2,$$

where the $\mathbb{Z}/2$-action on A_3 by conjugation is $\tau\sigma\tau^{-1} = \sigma^{-1}$. Similarly:

$$1 \to \mathbb{Z}/n \to D_n \to \mathbb{Z}/2 \to 1,$$

where $\mathbb{Z}/2 \simeq \{id, reflection\} \subset D_n$, so:

$$D_n \simeq \mathbb{Z}/n \rtimes \mathbb{Z}/2.$$

These are not direct products.

Remark 4.68. *If G is finite, $|G| = |H| \cdot |N|$, and $H \cap N = \{e\}$, then every coset of N contains a unique element of H. Assuming N is normal, we have a semidirect product by the proposition.*

Indeed: consider the homomorphism

$$H \to G/N,$$
$$h \mapsto hN$$

with

$$H \hookrightarrow G \twoheadrightarrow G/N.$$

This map has ker $= H \cap N = \{e\}$, so it is injective. Since $|H| = |G/N|$, it is also bijective. Alternatively, if $n_1h_1 = n_2h_2$, then $n_2^{-1}n_1 = h_2h_1^{-1} \in H \cap N = \{e\}$,

so $n_1 = n_2$ and $h_1 = h_2$. Thus, the products $n \cdot h, n \in N, h \in H$, are distinct, and every element of G has exactly one such expression. Hence, $|G| = |N||H|$.

Finally: if both N and H are normal subgroups of G, and every element of G can be uniquely expressed as $g = n \cdot h, n \in N, h \in H$ ($\iff$ every coset of one subgroup contains a unique elements of the other subgroup) then $G \simeq N \times H$ (ie. the semidirect product is actually a direct product).

This is because cosets intersect in a single element: $nH \cap Nh = \{nh\}$. Since H and N are normal, we have the following:

$$(n_1h_1)(n_2h_2) \in Nh_1 \cdot Nh_2 = Nh_1h_2$$

and

$$(n_1h_1)(n_2h_2) \in n_1H \cdot h_2H = n_1n_2H.$$

Thus,

$$(n_1h_1)(n_2h_2) \in n_1n_2H \cap Nh_1h_2,$$

which implies

$$(n_1h_1)(n_2h_2) = (n_1n_2)(h_1h_2),$$

showing that the map $N \times H \to G$, $(n, h) \mapsto nh$, is a group isomorphism.

Corollary 4.69. *If G is finite, $N, H \subset G$ are normal subgroups, $N \cap H = \{e\}$, and $|G| = |H| \cdot |N|$, then $G \simeq N \times H$.*

Remark 4.70. *The condition $N \cap H = \{e\}$ is, for instance, automatic if $\gcd(|N|, |H|) = 1$ (since $N \cap H$ is a subgroup of both N and H, so its order divides both $|N|$ and $|H|$).*

Returning to a group G of order 15, the Sylow theorems imply that G has unique subgroups H and K of orders 3 and 5, respectively, which are normal (uniqueness implies $gHg^{-1} = H$ and $gKg^{-1} = K$). Since $3 \cdot 5 = 15$ and $\gcd(3, 5) = 1$, the criterion holds, and so $G \simeq H \times K \simeq \mathbb{Z}/3 \times \mathbb{Z}/5 \simeq \mathbb{Z}/15$.

Example 4.71. *Another example: Consider groups of order 21. The Sylow theorems give the existence of subgroups H of order 3 and K of order 7. Also, the number of conjugate subgroups of each of these is:*

- $S_7 \equiv 1 \pmod 7$, *so* $S_7 = 1$.
- $S_3 \equiv 1 \pmod 3$, *and* $S_3 \mid 7$, *so S_3 could be either 1 or 7.*

If $S_3 = S_7 = 1$, then H and K are normal (since they are equal to their conjugates), and the above criterion implies that $G \simeq H \times K \simeq \mathbb{Z}/3 \times \mathbb{Z}/7 \simeq \mathbb{Z}/21$.

Otherwise, if $S_3 = 7$, then K is normal, but H is not, and we have a semidirect product $K \rtimes H$. Let x be a generator of $K \simeq \mathbb{Z}/7$ and y a generator of $H \simeq \mathbb{Z}/3$. Then $x^7 = y^3 = e$, and every element of G can be uniquely expressed as $x^a y^b$ with $0 \leq a \leq 6$ and $0 \leq b \leq 2$. To determine the group structure, we need to

know how $y \cdot x$ behaves. Since K is normal, $yx \in yK = Ky$, so $yx = x^\alpha y$ for some $0 \leq \alpha \leq 6$. Therefore, $yxy^{-1} = x^\alpha$, which fully determines the group law.

Furthermore, investigation shows that there exists a unique non-abelian group of order 21 up to isomorphism. The best way to prove existence is to construct it explicitly, for example, as a subgroup of S_7 or another group.

Next, we'll look at the proof of the Sylow theorems. For now, a couple comments:

Recall that for any $g \in G$, the order of g divides $|G|$, but the converse does not always hold. Specifically, in general, $k \mid |G|$ does not imply the existence of $g \in G$ such that $\text{ord}(g) = k$.

A corollary of Sylow's first theorem (the existence of Sylow p-subgroups) is that the converse does hold for primes.

Corollary 4.72. *If $p \mid |G|$ and p is prime, then G contains an element of order p.*

Proof. Let $H \subset G$ be a Sylow p-subgroup, and let $g \in H$ such that $g \neq e$. Since the order of g divides $|H| = p^e$, it follows that the order of g is p^k for some $1 \leq k \leq e$. Therefore, $g^{p^{k-1}}$ has order p. □

For a p-group ($|G| = p^n$), Sylow tells us essentially nothing! Specifically, a Sylow p-subgroup has p^n elements, and the only such subgroup is G itself. Thus, in the Sylow approach to classification, p-groups are the hardest to classify. In fact, the number of different p-groups grows rapidly with exponent n.

Example 4.73. *For $p = 2$, there exists:*

- *1 group of order 2,*
- *2 groups of order 4,*
- *5 groups of order 8,*
- *14 groups of order 16,*
- *51 groups of order 32.*

Another corollary of Sylow's first theorem:

Corollary 4.74. *If $p \mid |G|$ and p is prime, then G contains an element of order p.*

Proof. Let $H \subset G$ be a Sylow p-subgroup, and let $g \in H$ such that $g \neq e$. Since the order of g divides $|H| = p^e$, it follows that the order of g is p^k for some $1 \leq k \leq e$. Therefore, $g^{p^{k-1}}$ has order p. □

4.9 Proofs of Sylow Theorems

The first two theorems are proved by studying the action of G on its subsets by left multiplication.

The proof of Sylow's first theorem was two lemmas:

The proof of Sylow's first theorem relies on two lemmas:

Lemma 4.75. *Given $n = p^e m$ with $p \nmid m$, we have $p \nmid \binom{n}{p^e}$.*

Proof.

$$\binom{n}{p^e} = \frac{n(n-1)\cdots(n-p^e+1)}{p^e(p^e+1)\cdots 1} = \prod_{k=0}^{p^e-1} \frac{p^e m - k}{p^e - k}.$$

The highest power of p dividing $p^e m - k$ or $p^e - k$ is exactly the highest power of p dividing k (considering it modulo p^e). Hence, the numerator and denominator each contain some powers of p in their prime factorizations, and the end result contains no powers of p. □

Lemma 4.76. *Let $U \subset G$ be any subset, and consider the action of G on $P(G)$, the set of all subsets of G, by left multiplication. Then the stabilizer of $[U] \in P(G)$, $Stab([U]) = \{g \in G \mid gU = U\}$, satisfies $|Stab(U)|$ divides $|U|$.*

Proof. Let $H = \mathrm{Stab}(U)$. Then H acts on U by left multiplication ($hU = U$ for all $h \in H$), so U is a union of orbits $\mathcal{O}_u = \{hu \mid h \in H\} = Hu$ for various $u \in U$. Each orbit is a (right) coset of H, and has size $|\mathcal{O}_u| = |H|$. Since U is a union of such orbits, we conclude that $|H|$ divides $|U|$. □

Now, we can prove Sylow's first theorem (the existence of Sylow subgroups).

Proof. Let $S = \{U \in P(G) \mid |U| = p^e\}$ be the set of all subsets of G with p^e elements. Consider the action of G on S by left multiplication, $U \mapsto gU$, and partition S into orbits for this action. By Lemma 1, $p \nmid |S|$, so there exists an orbit $\mathcal{O}_U \subset S$ such that $p \nmid |\mathcal{O}_U|$. Since p^e divides $|G| = |\mathcal{O}_U||\mathrm{Stab}(U)|$, we find that p^e divides $|\mathrm{Stab}(U)|$. But by Lemma 2, $|\mathrm{Stab}(U)|$ divides $|U| = p^e$, so $|\mathrm{Stab}(U)| = p^e$. This shows that $\mathrm{Stab}(U)$ is a Sylow p-subgroup, and in fact, U is a right coset of $\mathrm{Stab}(U)$. □

Next, we prove Sylow's second theorem:

Theorem 4.77 (Sylow's Second Theorem)**.** *If $H \subset G$ is a Sylow p-subgroup and $K \subset G$ is any p-subgroup, then there exists a conjugate $H' = gHg^{-1}$ such that $K \subset H'$ (for $|K| = p^e$, this says that all Sylow p-subgroups are conjugate).*

Proof. Let C be the set of left cosets of H, and consider the action of G on C by left multiplication. This action is transitive, i.e., there is only one orbit, since $p \nmid |C| = \frac{|G|}{p^e} = m_j$. There exists $c_0 \in C$, namely $c_0 = [H]$, such that $\text{Stab}(c_0) = H$. Any G-action on a set with these properties works in the same way. Now, restrict the action of G on C to a p-subgroup K. The K-action on C has orbits whose sizes divide $|K|$, hence each orbit has a size that is a power of p.

Since $p \nmid |C|$, there is at least one fixed point, i.e., there exists $c \in C$ such that $k \cdot c = c$ for all $k \in K$. Thus, $K \subset \text{Stab}(c) = H'$, which is conjugate to $\text{Stab}(c_0) = H$ because c, c_0 belong to the same orbit of G. Specifically, if the coset gH is fixed by K, i.e., $kgH = gH$ for all $k \in K$, then there exists $k \in K$ such that $g^{-1}kgH = H$. This implies $g^{-1}kg \in H$, so $k \in gHg^{-1}$, and thus $K \subset gHg^{-1}$. □

Before we proceed with the proof of the third theorem, we need to discuss normalizers and conjugate subgroups.

Problem 4.78. *Given a group G and a subgroup H, what is the largest subgroup $K \subset G$ such that H is normal inside K?*

Observe: The issue is that the condition $gHg^{-1} = H$ may not hold for all $g \in G$, but it needs to hold for all $g \in K$.

Definition 4.79. *The **normalizer** of a subgroup $H \subset G$ is $N(H) = \{g \in G \mid gHg^{-1} = H\}$. This is a subgroup of G, and for subgroups $H \subset K \subset G$, H is normal in K if and only if $K \subset N(H)$.*

Example 4.80. *Let $G = S_3$:*

- *$H = \{id, \sigma = (123), \sigma^2\} = A_3 \subset S_3$, so $N(H) = G$ (since H is normal in G, even though for transpositions g, we have $g\sigma g^{-1} = \sigma^2 \neq \sigma$).*
- *$H = \{id, \tau\} \simeq \mathbb{Z}/2 \subset S_3$ for τ a transposition, so $N(H) = H$ (since $gHg^{-1} = H$ implies $g \in \{id, \tau\}$).*

The normalizer measures how close H is to being normal in G. If H is normal, then $N(H) = G$.

The group G acts by conjugation on the set of all its subgroups. The orbit of H is the set of its conjugate subgroups $gHg^{-1} \subset G$. If H is normal, then $\mathcal{O}_H = \{H\}$. The stabilizer of H is $\{g \in G \mid gHg^{-1} = H\} = N(H)$, so by the orbit-stabilizer theorem, the size of the orbit is $|\mathcal{O}_H| = |G/N(H)|$, and the set of subgroups conjugate to H corresponds to the cosets of $N(H)$.

Proposition 4.81. *The number of subgroups conjugate to H in G is $|G/N(H)|$.*

Now, let's prove Sylow's third theorem (the number of Sylow p-subgroups equals s_p where s_p divides m and $s_p \equiv 1 \pmod{p}$).

Proof. Consider the action of G on the set of Sylow p-subgroups by conjugation. By the second theorem, this action is transitive (all Sylow p-subgroups are conjugate), and if $H \subset G$ is any Sylow p-subgroup, the stabilizer is $\{g \in G \mid gHg^{-1} = H\} = N(H)$, so the size of the orbit is $s_p = |\mathcal{O}_H| = \frac{|G|}{|N(H)|}$.

Since $H \subset N(H) \subset G$ and $|H| = p^e$, we know that p^e divides $|N(H)|$, so $s_p = \frac{|G|}{|N(H)|} = \frac{|G|}{p^e} = m$.

Next, restrict the conjugation action of G on the set of all Sylow p-subgroups to H. Observe that H itself is fixed under conjugation ($hHh^{-1} = H$ for all $h \in H$), so this gives an orbit of size 1. We claim that this is the only orbit.

Indeed, if H' is a Sylow p-subgroup such that $hH'h^{-1} = H'$ for all $h \in H$, then $H \subset N(H')$. But $|N(H')|$ is a multiple of $|H'| = p^e$, a divisor of $|G| = p^e m$. Therefore, H and H' are Sylow p-subgroups of $N(H')$, and since H' is normal in $N(H')$, we conclude that $H = H'$.

Thus, the only orbit of size 1 under the action of H on the set of Sylow p-subgroups is $\{H\}$ itself. Since the size of an orbit of an H-action divides $|H| = p^e$, all other orbits must have sizes divisible by p. We conclude that $s_p = \#\{\text{Sylow } p\text{-subgroups}\} \equiv 1 \pmod{p}$. □

One more example to show that things can get complicated quickly:

Example 4.82. *Let's try to classify groups of order 12. If $|G| = 12$, then by Sylow's theorems, we have:*

- *A subgroup $H \subset G$, with $|H| = 4$. The number of such subgroups is $s_2 \in \{1, 3\}$ ($s_2 \mid 3$, $s_2 \equiv 1 \pmod 2$).*
- *A subgroup $K \subset G$, with $|K| = 3$. The number of such subgroups is $s_3 \in \{1, 4\}$ ($s_3 \mid 4$, $s_3 \equiv 1 \pmod 3$).*

At least one of these subgroups must be normal. Indeed, if $s_3 = 4$, then the nontrivial elements of k_1, k_2, k_3, k_4 all have order 3, and $k_i \cap k_j = \{e\}$ (since the order divides 3 and is less than 3). Hence, we have 8 elements of order 3. This leaves at most 4 elements of order in $\{1, 2, 4\}$, which forces $s_2 = 1$ and implies that H is normal.

If both H and K are normal, then $G \simeq H \rtimes K$ (using $|G| = |H| \cdot |K|$ and $H \cap K = \{e\}$), and so G is abelian. Specifically, G is isomorphic to one of the following:

$$\mathbb{Z}/4 \times \mathbb{Z}/3 \simeq \mathbb{Z}/12 \quad \text{or} \quad (\mathbb{Z}/2 \times \mathbb{Z}/2) \times \mathbb{Z}/3 \simeq \mathbb{Z}/2 \times \mathbb{Z}/6.$$

If H is normal but K is not, consider the action of G on $\{k_1, k_2, k_3, k_4\}$ by conjugation. Conjugation by a nontrivial element of K_1 maps K_1 to itself but does not fix any of the other 3 subgroups. Recall that the stabilizer of K_i is

$$N(K_i) = \{g \in G \mid gK_ig^{-1} = K_i\}.$$

By the orbit-stabilizer theorem, we have

$$|N(K_i)| = \frac{|G|}{s_3} = \frac{12}{4} = 3,$$

so $N(K_i) = K_i$. Thus, a nontrivial element of K_1 acts on $\{K_1, K_2, K_3, K_4\}$ by a 3-cycle, permuting $\{K_2, K_3, K_4\}$, and similarly for the other subgroups. Therefore, the action of G on $\{K_1, \ldots, K_4\}$ gives a homomorphism $\varphi : G \to S_4$, where $y_i \mapsto$ 3-cycles. This implies that $Im(\varphi) \supset A_4$, and hence $Im(\varphi) = A_4$. Therefore, $G \simeq A_4$.

If K is normal but H is not, there are two subcases: $H \simeq \mathbb{Z}/4$ or $H \simeq \mathbb{Z}/2 \times \mathbb{Z}/2$.

- *If $H \simeq \mathbb{Z}/4$, let $x \in H$ be a generator, and let $K = \{e, y, y^2\}$. Then $G \simeq K \rtimes H$, determined by the conjugation action of H on K. Specifically, we need to know $xyx^{-1} \in K$. We cannot have $xyx^{-1} = e$ (which would imply $y = e$) or $xyx^{-1} = y$ (which would imply that x and y commute, making G abelian). Therefore, we must have $xyx^{-1} = y^2 = y^{-1}$. In this case, G is generated by x and y, with relations $x^4 = y^3 = e$ and $xy = y^2x$. This is a semidirect product $\mathbb{Z}/3 \rtimes \mathbb{Z}/4$, where $\mathbb{Z}/4$ acts on the normal subgroup $\mathbb{Z}/3$ by the automorphism $\mathbb{Z}/4 \to Aut(\mathbb{Z}/3) = \{\pm id\}$, with the action $k \mapsto (-1)^k$.*
- *If $H \simeq \mathbb{Z}/2 \times \mathbb{Z}/2$, consider the conjugation action $H \xrightarrow{\varphi} Aut(K) \simeq \mathbb{Z}/2$, which must have kernel $Ker(\varphi) \simeq \mathbb{Z}/2$. Denote its generator by z, and let $x \in H$ such that x and z generate H, with y a generator of K. Then G is generated by x, y, z, with relations $x^2 = z^2 = y^3 = e$, $xz = zx$, $zy = yz$, and $xy = y^2x$. This gives $G \simeq D_6$, the dihedral group of order 6. The subgroup generated by y and z is isomorphic to $\mathbb{Z}/6$ and normal in G. In this case, y corresponds to a rotation by $\frac{2\pi}{3}$, z corresponds to a rotation by π, and x is any reflection.*

Thus, there are 5 isomorphism classes of groups of order 12:

$$\mathbb{Z}/12, \quad \mathbb{Z}/2 \rtimes \mathbb{Z}/6, \quad A_4, \quad \mathbb{Z}/3 \rtimes \mathbb{Z}/4, \quad D_6.$$

4.10 Generators, Presentations, and Cayley Graph

Definition 4.83. *The **free group** F_n on n generators $a_1, \ldots, a_n$ consists of elements that are **reduced words** of the form $a_{i_1}^{m_1} a_{i_2}^{m_2} \ldots a_{i_k}^{m_k}$ for $k \geq 0$ (with the empty word being denoted by e), where $i_1, i_2, \ldots, i_k \in \{1, \ldots, n\}$, $i_j \neq i_{j+1}$, and $m_1, m_2, \ldots, m_k \in \mathbb{Z} - \{0\}$.*

For non-reduced words, we apply the following reduction rules:

- *If $i_j = i_{j+1}$, combine $a_i^m a_i^{m'}$ into $a_i^{m+m'}$.*
- *If an exponent is zero, remove a_i^0.*

This process is repeated until the word is fully reduced.

The free group F_n is the largest group with n generators. Any other group generated by n elements is isomorphic to a quotient of F_n. Specifically, if G is generated by $g_1, \ldots, g_n \in G$, we define a homomorphism $\varphi : F_n \twoheadrightarrow G$ by setting $a_i \mapsto g_i$ for all i. Consequently, a word in F_n of the form $a_{i_1}^{m_1} a_{i_2}^{m_2} \ldots a_{i_k}^{m_k}$ maps to the corresponding word in G, i.e., $\prod a_{ij}^{m_j} \mapsto \prod g_{ij}^{m_j}$.

Definition 4.84. *A finitely generated group is said to be* ***finitely presented*** *if the kernel of φ is the smallest normal subgroup of F_n containing some finite set of relations $\{r_1, \ldots, r_k\} \subset F_n$, i.e., the subgroup generated by the r_j's and their conjugates $c^{-1} r_j c$.*

We then write $G \simeq \langle a_1, \ldots, a_n \mid r_1, \ldots, r_k \rangle$, where $G \simeq F_n / \langle \text{conjugates of } r_1, \ldots, r_k \rangle$.

Example 4.85.

- $\mathbb{Z}^n \simeq \langle a_1, \ldots, a_n \mid a_i a_j a_i^{-1} a_j^{-1} \text{ for all } i, j \rangle$.
- $S_3 \simeq \langle s_1, s_2 \mid s_1^2, s_2^2, (s_1 s_2)^3 \rangle$.

In practice, given generators $g_1, \ldots, g_n \in G$, it is often possible to find relations $r_1, \ldots, r_k$, i.e., words in the free group F_n, such that under the homomorphism $\varphi : F_n \to G$ defined by $a_i \mapsto g_i$, we have $r_j \mapsto e$. If these relations hold in G, then φ induces a surjective homomorphism $\langle a_1, \ldots, a_n \mid r_1, \ldots, r_k \rangle = F_n / \langle \text{conjugates of } r_j \rangle \twoheadrightarrow G$. This is an isomorphism once we have identified a complete set of relations among the g_i, i.e., when the relations $r_1, \ldots, r_k$ and their conjugates generate $\mathrm{Ker}(\varphi)$.

Problem 4.86. *How should one work with a group described by generators and relations?*

In some cases, we may already know what G is, while in others, we may not. Two useful concepts (among many others) are:

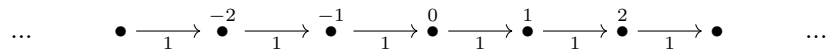
1. The Cayley Graph.
2. Normal Forms.

Definition 4.87. *Given generators $g_1, \ldots, g_n \in G$, the* ***Cayley graph*** *of G is constructed such that*

- *The vertices correspond to the elements of the group.*
- *Two vertices s and t are connected by an edge labeled g_i when $t = s g_i$.*

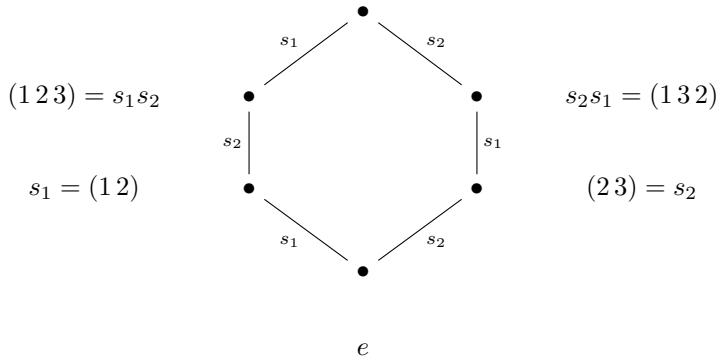
Remark 4.88. *Here, we are using right multiplication; one could use left multiplication instead.*

Example 4.89. *Consider $\mathbb{Z}$ with its usual generator 1. It's Cayley graph is given as follows:*



Example 4.90. *Consider S_3 with generators $s_1 = (1\,2)$, $s_2 = (2\,3)$ (note that*

$s_i^{-1} = s_i$, *so edges are undirected). It's Cayley graph is given as follows:*



The fact that this graph closes up shows the relation $s_1 s_2 s_1 = s_2 s_1 s_2$ *(which implies* $(s_1 s_2)^3 = e$*). Since any word in* s_1, s_2 *with the relations* $s_1^2 = s_2^2 = e$ *can be reduced, one can use this to verify that* $S_3 \simeq \langle s_1, s_2 \mid s_1^2, s_2^2, (s_1 s_2)^3 \rangle$.

Example 4.91. *Consider* S_4 *with generators* $s_i = (i\, i+1)$ *for* $i \leq 3$*. This generates a permutahedron, where faces are square relations* $s_1 s_3 = s_3 s_1$ *and hexagonal relations* $s_1 s_2 s_1 = s_2 s_1 s_2$ *and* $s_2 s_3 s_2 = s_3 s_2 s_3$.

More generally, S_n has the following presentation:

$$S_n = \langle s_1, \ldots, s_{n-1} \mid s_i^2 = 1 \text{ for all } i, s_i s_j = s_j s_i \text{ for } |i-j| \geq 2, s_i s_{i+1} s_i = s_{i+1} s_i s_{i+1} \rangle.$$

Proposition 4.92. *G acts on its Cayley graph by "left multiplication": vertices $s \mapsto gs$, edges $(s \mapsto sg_i) \mapsto (gs \mapsto gsg_i)$.*

This action is transitive on vertices (and on edges with a given label g_i), which makes the graph very symmetric.

Definition 4.93. *The* ***word length*** *of an element $g \in G$ is the shortest distance from the identity element e to g in the Cayley graph.*

For infinite groups, we can inquire about the growth rate of G. Given a set of generators g_i, how does the number of elements represented by words of length $\leq N$ grow with N? Does it grow polynomially or exponentially?

Even if we change our set of generators to some other set g'_j, the word length of a given element changes by a bounded factor only. The bound is determined by the word lengths of the new generators in terms of the old ones and vice versa. Therefore, the exponential or polynomial nature of the growth is independent of the set of generators.

Example 4.94. *Abelian groups have polynomial growth, while free groups have exponential growth.*

4.11 Braids

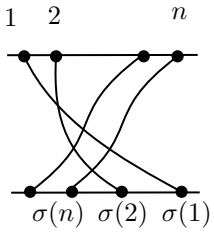
For finite groups, the Cayley graph is finite, and growth isn't relevant. However, questions about word length remain interesting!

In S_n, where $\{s_i = (i\, i+1)\}$, the largest element is

$$\begin{bmatrix} 1 & 2 & \dots & n \\ J & & & J \\ n & n-1 & \dots & 1 \end{bmatrix}$$

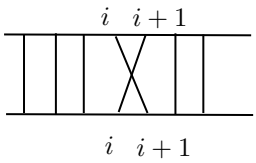
with a word length of $\frac{n(n-1)}{2}$. The word length of $\sigma \in S_n$ is the number of inversions, i.e., the pairs (i, j) such that $i < j$ and $\sigma(i) > \sigma(j)$.

This is best understood by representing permutations as diagrams:



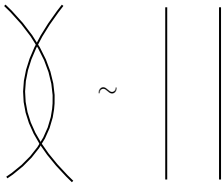
where

- **Composition:** Stack diagrams.
- The expression in terms of s_i comes from decomposing the diagram into layers with a single crossing:

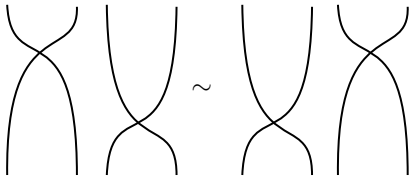


$= s_i$

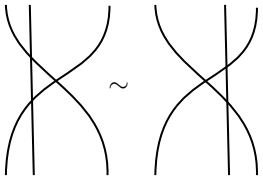
- Presentation of $S_n \iff$ Any two diagrams for σ are related by three braid relations:



1. $s_i^2 = e$



2. $s_i s_j = s_j s_i \qquad |i-j| \geq 2$



3. $s_i s_{i+1} s_i = s_{i+1} s_i s_{i+1}$

The word length, defined as the number of inversions, becomes clearer. We can even list all the shortest words that represent a given permutation. Namely, $\sigma \in S_n$ has a shortest word ending with $s_i \iff w(\sigma s_i) < w(\sigma) \iff \sigma(i+1) < \sigma(i)$. We call the set of such i the "ending set" of σ. Then, for each $i \in$ the ending set, repeat the process for $\sigma s_i^{-1} = \sigma s_i$.

For each $\sigma \in S_n$, we can find a preferred expression of σ as a word in $s_1, \ldots, s_n$ by choosing, at each step, the smallest i such that $\sigma(i+1) < \sigma(i)$ to end the word. This provides a normal form for elements of S_n (a preferred word representing each element) and thus a solution to the word problem: When do two words represent the same element? (i.e., when does a word represent $e \in G$?)

For S_n or other groups where it isn't well understood how to calculate elements (e.g., groups of matrices), we don't need fancy algorithms or normal forms to solve the word problem. However, in many groups, this is all we have!

Definition 4.95. *The **braid group** is defined as*

$$B_n = \langle s_1, \ldots, s_{n-1} \mid s_i s_j = s_j s_i \, \forall |i-j| \geq 2,\ s_i s_{i+1} s_i = s_{i+1} s_i s_{i+1} \rangle,$$

but $s_i^2 \neq 1$.

Now let's talk about Markov's theorem:

Theorem 4.96 (Markov's Theorem)**.** *Every knot or link in $\mathbb{R}^3$ can be represented as the closure of a braid. Two braids have isotopic closures if and only if they are related by a sequence of moves of two types:*

- *Conjugation in B_n: For $\sigma \in B_n$, we have $\sigma \sim g\sigma g^{-1} \in B_n$ for all $g \in B_n$.*
- *Stabilization: $B_n \simeq \langle s_1, \ldots, s_{n-1} \rangle \subset B_{n+1}$, and $\sigma \in B_n$ is related to $\sigma S_n^{\pm 1} \in B_{n+1}$.*

Braids play an important role in knot theory, and their algorithmics are similar to those of S_n. Permutation braids are defined similarly, and any two strands can cross at most once. These form a finite set, in bijection with S_n.

Let Δ be the longest permutation braid. Since its shortest word can start or end with any s_i, the word Δ is still a permutation braid, and it conjugates $s_i \iff s_{n-i}$. Thus, any element of B_n can be written as $g = \Delta^{-k} P_1 P_2 \cdots P_r$, where each P_j is a permutation braid. Additionally, "moving to the left everything that can be" implies we can find an expression such that

$$\{\text{ending set of } P_j\} \supset \{\text{starting set of } P_{j+1}\} \quad \forall j.$$

In other words, any attempt to add an initial letter of a shortest word of P_{j+1} to the end of P_j would cause it to no longer be a permutation braid.

This result leads to the Garside normal form and provides a solution to the word problem in B_n.

One more example:

Example 4.97. *Consider $SL_2(\mathbb{Z})$ and $PSL_2(\mathbb{Z}) = SL_2(\mathbb{Z})/\{\pm I\}$. $SL_2(\mathbb{Z})$ is generated by $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ and $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$.*

Proof. Given $M = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}_2(\mathbb{Z})$, we wish to express it in terms of the generators S and T.

- If $c = 0$, then $a, d = \pm 1$. In this case, M is either $\begin{bmatrix} 1 & n \\ 0 & 1 \end{bmatrix} = T^n$ or $\begin{bmatrix} -1 & -n \\ 0 & -1 \end{bmatrix} = S^2 T^n$.
- Now assume $c \neq 0$, and apply the following algorithm to modify M:
  - If $|a| \geq |c|$, use Euclidean division to write $a = nc + r$, where $|r| < |c|$. Then
    $$T^{-n} M = \begin{bmatrix} -1 & -n \\ 0 & -1 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a - nc & b - nd \\ c & d \end{bmatrix},$$
    which decreases $\max(|a|, |c|)$.
  - If $|a| < |c|$, apply $S^{-1} M = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} c & d \\ -a & -b \end{bmatrix}$, which brings us back to the case $|a| > |c|$.

After finitely many steps, we find that the product of M with some word in S and T has $c = 0$ and $|a| = 1$, hence it is either T^n or $S^2 T^n$. □

There is an alternative, geometric proof based on the fact that $\mathrm{PSL}_2(\mathbb{Z})$ acts on the upper half-plane $\mathbb{H} = \{z \in \mathbb{C} \mid \mathrm{Im}(z) > 0\}$ by the Möbius transformation

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} : z \mapsto \frac{az+b}{cz+d}.$$

Here, S acts by $z \mapsto -\frac{1}{z}$ and T by $z \mapsto z+1$.

The region $\Delta = \{z \in \mathbb{C} \mid |z| \geq 1, |\mathrm{Re}(z)| \leq \frac{1}{2}\}$ is a **fundamental domain** of this action, in the sense that Δ and its images under $\mathrm{PSL}_2(\mathbb{Z})$ tile $\mathbb{H}$. Since the regions immediately adjacent to Δ are $T^{\pm 1}(\Delta)$ and $S(\Delta)$, the regions adjacent to $g(\Delta)$ are $gT^{\pm 1}(\Delta)$ and $gS(\Delta)$. The structure of the tiling exactly reproduces the Cayley graph of $\mathrm{PSL}_2(\mathbb{Z})$ with generators S and T, and the fact that all regions can be reached from Δ in finitely many steps implies that S and T generate $\mathrm{PSL}_2(\mathbb{Z})$.

There also exist other generators: Instead of S and T, we can also use:

- $S = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ and $R = ST = \begin{bmatrix} 0 & -1 \\ 1 & 1 \end{bmatrix}$, which have finite order: $S^4 = R^6 = I$.
- $T = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ and $T' = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} = (TST)^{-1}$, which are conjugates: $T' = STS^{-1}$.
- The images of these matrices in $\mathrm{PSL}_2(\mathbb{Z})$ also generate $\mathrm{PSL}_2(\mathbb{Z})$.

Theorem 4.98.

$$PSL_2(\mathbb{Z}) \simeq \langle S, R \mid S^2, R^3 \rangle.$$

Proof. Since $S^2 = -I$ and $R^3 = -I$, S and R have orders 2 and 3 in $\mathrm{PSL}_2(\mathbb{Z})$. These relations $S^2 = R^3 = e$ reduce any word in $S^{\pm 1}, R^{\pm 1}$ to the form $\ldots SR^{\pm 1}SR^{\pm 1}SR^{\pm 1}\ldots$. (The word can start and end with either S or $R^{\pm 1}$.) Mapping $F_2 = \langle s, r \rangle$ to $\mathrm{PSL}_2(\mathbb{Z})$ by $r \mapsto R, s \mapsto S$ induces a surjective homomorphism

$$\langle s, r \mid s^2, r^3 \rangle = F_2 / \langle \text{conjugations of } s^2, r^3 \rangle \twoheadrightarrow \mathrm{PSL}_2(\mathbb{Z}),$$

and the kernel consists of elements whose corresponding expression in S and R equals e in $\mathrm{PSL}_2(\mathbb{Z})$, i.e., $\pm I$ in $\mathrm{SL}_2(\mathbb{Z})$.

Observe that $S \neq e$ and $R^{\pm 1} \neq e$. If some longer word w in S and $R^{\pm 1}$'s (alternating between these) simplifies to $e \in \mathrm{PSL}_2(\mathbb{Z})$, we have:

- If it starts and ends with S, conjugating by S yields a shorter word w': since $S^2 = e$, $Sw'S = e \iff w' = e$.
- If it starts with $R^{\pm 1}$, conjugating gives another word that doesn't simplify to e: $R^{\pm 1}w' = e \iff w'R^{\pm 1} = e$.

Iterating this process, we eventually obtain a word of the form $SR^{\pm 1} \dots SR^{\pm 1} = \pm I$. However, $SR = -T = -\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ and $SR^{-1} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$. A product of these matrices cannot simplify to $\pm I$, as multiplying matrices with non-negative entries by $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ or $\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$ increases the sum of the entries.

Thus, no word in $SR^{\pm 1}$ simplifies to e, and we are done. □

This presentation can be rewritten in terms of other generators:

$$\mathrm{PSL}_2(\mathbb{Z}) \simeq \langle S, T \mid S^2, (ST)^3 \rangle \simeq \langle T, T' \mid (TT')^3 = e, TT'T = T'TT' \rangle,$$

and

$$\mathrm{SL}_2(\mathbb{Z}) = \langle T, T' \mid (TT')^6 = 1, TT'T = T'TT' \rangle.$$

(For constructing the two strands of the discussion, the center of the braid group $B_3 = \langle s_1, s_2 \mid s_1 s_2 s_1 = s_2 s_1 s_2 \rangle$ is generated by $\Delta^2 = (s_1 s_2)^3$, and $\mathbb{Z} \cong \langle \Delta^2 \rangle$. Thus, mapping $s_1 \mapsto T, s_2 \mapsto T'$ gives a sequence $1 \to \mathbb{Z} = \langle \Delta^2 \rangle \hookrightarrow B_3 \twoheadrightarrow \mathrm{PSL}_2(\mathbb{Z}) \to 1$.)

5 Representation Theory

5.1 Representations

Representation theory is the study of group actions on vector spaces, i.e., homomorphisms $G \to \mathrm{GL}(V)$ (typically with $k = \mathbb{C}$). Historically, groups first arose as geometric symmetries, and in the 19th century, groups were primarily seen as subgroups of $\mathrm{GL}(n)$, rather than abstract entities. The modern viewpoint divides this into two main components: the study of groups themselves (which we've explored) and the interpretation of an abstract group G as a subgroup of $\mathrm{GL}(n)$ (which we will now consider). While we focus on representations of finite groups, this problem is also significant for discrete finite groups (e.g., $\mathrm{SL}_2(\mathbb{Z})$, braid groups, etc.), and continuous groups (Lie groups such as S^1, $\mathrm{SO}(3)$, etc.).

Definition 5.1. *A **representation** of a group G is a vector space V with an action of G on V by linear operators, i.e., $G \times V \to V$ such that for all $g \in G$, the map $g : V \to V$ is linear.*

Equivalently, a representation is a homomorphism $\rho : G \to \mathrm{GL}(V)$, where $\mathrm{GL}(V)$ is the group of invertible linear operators on V.

Definition 5.2. *A **subrepresentation** of a representation is a subspace $W \subset V$ that is invariant under the action of G, i.e., $gW = W$ for all $g \in G$.*

Definition 5.3. *A representation is **irreducible** if it has no nontrivial subrepresentations.*

Example 5.4. *If $G \cong \mathbb{Z}/n$ is a cyclic group, then a representation of G is a vector space V together with a map $\varphi = \rho(1) : V \to V$ such that $\varphi^n = id_V$.*

Lemma 5.5. *If V is a finite-dimensional $\mathbb{C}$-vector space and $\varphi : V \to V$ has finite order, i.e., $\varphi^n = id$, then φ is diagonalizable.*

Proof. This follows because the minimal polynomial of φ divides $\varphi^n - 1$, which factors as a product of linear factors with distinct roots. Specifically, over $\mathbb{C}$, we have $\varphi^n - 1 = 0$ factoring as $\prod_k (\varphi - \lambda_k) = 0$, where $\lambda_k = e^{\frac{2\pi i k}{n}}$ are the nth roots of unity. Therefore, the eigenvalues of φ are these nth roots of unity, and we obtain a direct sum decomposition of V into eigenspaces. Since each eigenspace is invariant under φ, φ is diagonalizable. □

Returning to the example, the invariant subspaces of $\varphi = \rho(1)$ are subrepresentations, and V splits into a direct sum of 1-dimensional irreducible representations. These correspond to homomorphisms $\mathbb{Z}/n \to \mathbb{C}^\times = \mathrm{GL}_1(\mathbb{C})$, with $1 \mapsto \lambda = e^{\frac{2\pi i k}{n}}$.

Now, if V is a $\mathbb{C}$-representation of a finite abelian group $G \cong \mathbb{Z}/m_1 \times \mathbb{Z}/m_2 \times \cdots \times \mathbb{Z}/m_r$, the G-action is equivalent to the data of $\varphi_1, \ldots, \varphi_r : V \to V$ such that $\varphi_i^{m_i} = \mathrm{id}_V$ and the φ_i pairwise commute, i.e., $\varphi_i \varphi_j = \varphi_j \varphi_i$. Each φ_i is diagonalizable by the lemma, and commuting diagonalizable operators are

simultaneously diagonalizable. By induction on r, this shows that V splits into a direct sum of 1-dimensional subrepresentations, each corresponding to a homomorphism $G \to \mathrm{GL}_1(\mathbb{C}) \cong \mathbb{C}^\times$.

For a finite abelian group G, define its **dual** as $\hat{G} = \mathrm{Hom}(G, \mathbb{C}^\times)$. This is an abelian group under pointwise multiplication. If $\rho, \rho' \in \mathrm{Hom}(G, \mathbb{C}^\times)$, their product $\rho\rho'$ is also a homomorphism, defined by

$$(\rho\rho')(g) = \rho(g)\rho'(g) \quad \text{for all } g \in G.$$

For example, if $G \cong \mathbb{Z}/n$, then $\hat{G} \cong \mathbb{Z}/n$, with $\rho \mapsto \rho(1)$ being a map into the nth roots of unity, which forms a group isomorphic to $\mathbb{Z}/n$.

Similarly, if $G = \mathbb{Z}/m_1 \times \cdots \times \mathbb{Z}/m_r$, then $\hat{G} \cong \hat{\mathbb{Z}/m_1} \times \cdots \times \hat{\mathbb{Z}/m_r}$, and the homomorphisms are determined by the images of the generators of G, which are roots of unity in $\mathbb{C}^\times$. This completes the classification of complex representations of finite abelian groups!

Definition 5.6. *Given two representations V and W of G, a **homomorphism** of representations $\varphi : V \to W$ is a linear map that is equivariant, i.e., satisfies $\varphi(gv) = g\varphi(v)$ for all $v \in V$ and $g \in G$.*

The set of homomorphisms of representations, denoted by $\mathrm{Hom}_G(V, W)$, consists of G-equivariant linear maps (as opposed to all linear maps, which form $\mathrm{Hom}(V, W)$).

We can construct new representations from existing ones. In particular:

- If V, W are representations of G and $\varphi \in \mathrm{Hom}_G(V, W)$, then $\mathrm{Ker}(\varphi)$ and $\mathrm{Im}(\varphi)$ are preserved by G, hence they are subrepresentations of V and W.
- If $W \subset V$ is a subrepresentation, then V/W is also a representation, as G acts on cosets: $g(v + W) = gv + W$.
- If V and W are representations of G, then $V \oplus W$ is also a representation with $g(v, w) = (gv, gw)$.
- The tensor product $V \otimes W$ is a representation, where $g(v \otimes w) = gv \otimes gw$ (extended by linearity).
- $\mathrm{Hom}(V, W)$ (the space of all linear maps) can also be given the structure of a representation, but the action of g is given by conjugation: $g(\varphi)(v) = g\varphi(g^{-1}v)$.
- Specializing to $V^* = \mathrm{Hom}(V, k)$, where k has the trivial representation, the dual representation of V is V^* with $g(\ell) = \ell \circ g^{-1}$.

Theorem 5.7. *Let V be a representation of a finite group G (over $\mathbb{C}$ or any field of characteristic zero), and suppose $W \subset V$ is an invariant subspace (a subrepresentation). Then there exists another invariant subspace $U \subset V$ such that $V = U \oplus W$ as a direct sum of representations.*

Corollary 5.8. *Any finite-dimensional representation of a finite group decomposes into a direct sum of irreducible representations.*

To prove this, let's first prove a lemma.

Lemma 5.9. *If V is a $\mathbb{C}$-representation of a finite group G, then there exists a positive definite Hermitian inner product on V that is preserved by G, i.e., $H(gv, gw) = H(v, w)$ for all $v, w \in V$ and $g \in G$. In other words, all the linear operators $g : V \to V$ are unitary.*

Proof. Let H_0 be any Hermitian inner product on V, and define $H(v, w) = \frac{1}{|G|} \sum_{g \in G} H_0(gv, gw)$. This new inner product is Hermitian, positive definite, and invariant under the action of G. □

Now we proceed with the first proof of the theorem.

Proof. Equip V with a G-invariant Hermitian inner product H as in the lemma. If $g(W) = W$, then g is unitary, implying that $g(W^\perp) = W^\perp$. Therefore, $U = W^\perp$ is a complementary invariant subspace, and we have $V = U \oplus W$. □

Next, let's consider an alternate proof.

Proof. Choose any complementary subspace $U_0 \subset V$ such that $V = U_0 \oplus W$. Define the projection $\pi_0 : V \to W$ onto W with kernel U_0. Then define $\pi(v) = \frac{1}{|G|} \sum_{g \in G} g\pi_0(g^{-1}v) \in W$. The map π is a homomorphism of representations, and its kernel $U = \mathrm{Ker}(\pi)$ is an invariant subspace. Since $\pi|_W = \mathrm{id}$, it is surjective, and we conclude that $V = U \oplus W$. □

Remark 5.10. *The proof fails if char$(k) \neq 0$ (specifically, if char$(k) = p$ and $|G|$ is divisible by p), which is one reason why modular representations (representations over fields of positive characteristic) are more complicated. The proof also fails if G is infinite (and does not carry a finite invariant measure), as the averaging trick does not work in this case.*

Example 5.11. *If $G = \mathbb{Z}$ or $\mathbb{R}$ acts on $\mathbb{C}^2$ by $t \mapsto \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}$, then the first factor $\mathbb{C} \times 0$ is invariant under the action of G, but there is no complementary invariant subspace.*

5.2 Irreducibility and Representations of S_3

Goal: Given a group G, find its irreducible representations and describe how other representations decompose into irreducibles.

Theorem 5.12 (Schur's Lemma)**.** *Let V, W be irreducible representations of a group G, and let $\varphi : V \to W$ be a homomorphism of representations. Then, either $\varphi = 0$, or φ is an isomorphism.*

If $k = \mathbb{C}$, and V is an irreducible representation of G, then any homomorphism $\varphi : V \to V$ is a scalar multiple of the identity map, i.e., $\varphi = \lambda \cdot id$ for some $\lambda \in \mathbb{C}$.

Proof. Let $\varphi : V \to W$ be a homomorphism. The kernel of φ, denoted $\mathrm{Ker}(\varphi)$, is an invariant subspace of V, i.e., it is a subrepresentation. Since V is irreducible, we have two possibilities: either $\mathrm{Ker}(\varphi) = 0$ (in which case φ is injective), or $\mathrm{Ker}(\varphi) = V$ (in which case $\varphi = 0$).

Similarly, the image of φ, $\mathrm{Im}(\varphi)$, is an invariant subspace of W. Hence, $\mathrm{Im}(\varphi) = 0$ (if $\varphi = 0$) or $\mathrm{Im}(\varphi) = W$ (if φ is surjective). Therefore, we conclude that either $\varphi = 0$, or φ is an isomorphism.

For $k = \mathbb{C}$, suppose $\varphi : V \to V$ is a homomorphism. Then, φ has an eigenvalue λ, and the map $\varphi - \lambda I : V \to V$ has a nonzero kernel. Since V is irreducible, we deduce that $\varphi - \lambda I = 0$, implying $\varphi = \lambda I$.

□

Example 5.13. *Let V be an irreducible representation of G, and let $h \in Z(G)$ be an element of the center of G (i.e., h commutes with every $g \in G$). The action of h on V satisfies $h(gv) = gh(v)$ for all $g \in G$, so h is equivariant. By Schur's Lemma, $h \in Hom_G(V, V)$ must act by a scalar multiple of the identity, i.e., $h = \lambda \cdot id$ for some $\lambda \in \mathbb{C}$.*

In particular, if G is abelian and V is irreducible, then every element of G acts by a scalar multiple of the identity. This provides an alternate proof that irreducible representations of finite abelian groups are 1-dimensional.

Next, we consider the simplest non-abelian group S_3. We know that the trivial representation $U \simeq \mathbb{C}$ (where every element $\sigma \in S_3$ acts as the identity). Another 1-dimensional representation $U' \simeq \mathbb{C}$ corresponds to the **alternating representation**, where $\sigma \in S_3$ acts by $(-1)^\sigma$.

We also have the **permutation representation** $\simeq \mathbb{C}^3$ with basis e_1, e_2, e_3, where S_3 acts by permuting the basis vectors: $\sigma e_i \mapsto e_{\sigma(i)}$.

This representation has an invariant subspace, namely $\mathrm{span}(e_1 + e_2 + e_3)$, and we can find a complementary subrepresentation, specifically $V = \{(z_1, z_2, z_3) \in \mathbb{C}^3 \mid z_1 + z_2 + z_3 = 0\}$. This is the **standard representation** of S_3, with $\dim(V) = 2$, and it is irreducible.

Remark 5.14. *Similarly, for S_n, the two 1-dimensional representations are the trivial representation $U = \mathbb{C}$ and the alternating representation $U' = \mathbb{C}$ where σ acts by $(-1)^\sigma$. The permutation representation $\mathbb{C}^n$, where σ acts by permuting the basis vectors $e_i \mapsto e_{\sigma(i)}$, has an invariant subspace span$(e_1 + \cdots + e_n) \simeq U$, with a complementary subrepresentation $V = \{(z_1, \ldots, z_n) \in \mathbb{C}^n \mid \sum z_i = 0\}$. It turns out that V is irreducible, and it is the standard representation of S_n, with* $\dim(V) = n - 1$.

For S_3, however, this exhausts the list of irreducible representations (over $\mathbb{C}$). The group S_n has more irreducible representations. In fact, the number of irreducible representations of S_n corresponds to the number of partitions of n.

Proposition 5.15. *The irreducible representations of S_3 over $\mathbb{C}$ are U, U', and V. Hence, any representation of S_3 is isomorphic to a direct sum $U^{\oplus a} \oplus U'^{\oplus b} \oplus V^{\oplus c}$ for some unique $a, b, c \in \mathbb{N}$.*

Proof. Let W be any finite-dimensional representation of S_3. Restricting to the abelian subgroup $A_3 \simeq \mathbb{Z}/3 \subset S_3$, let $\tau \in S_3$ be a 3-cycle, and $\sigma \in S_3$ be a transposition. We have $\tau^3 = \sigma^2 = \mathrm{id}$ and $\sigma^{-1}\tau\sigma = \tau^2$. Restricting the representation to the subgroup generated by τ, W has a basis of eigenvectors (v_j), where $\tau(v_j) = \lambda_j v_j$ and $\lambda_j = e^{\frac{2\pi i k_j}{3}}$ are roots of unity.

Now consider how σ acts. If v is an eigenvector of τ with eigenvalue λ, then $\tau(\sigma v) = \lambda^2 \sigma(v)$. Thus, σ maps the λ-eigenspace of τ to its λ^2-eigenspace.

For each eigenvector v of τ, $\mathrm{span}(v, \sigma v)$ is an invariant subspace. Therefore, irreducible representations have $\dim \leq 2$.

Now, let W be irreducible. Consider the case where $v \in W$ is an eigenvector of τ.

Case 1: $\lambda = 1$. If $\tau(v) = v$, then $\tau(\sigma v) = \sigma v$. If σv is linearly independent of v, then $w = v + \sigma v$ is an invariant subspace, which contradicts the irreducibility of W. Thus, σv must be a scalar multiple of v, and since $\sigma^2 = \mathrm{id}$, we have $\sigma(v) = \pm v$. Therefore, W is isomorphic to either U or U', depending on whether $\sigma(v) = v$ or $\sigma(v) = -v$.

Case 2: $\lambda = e^{\pm \frac{2\pi i}{3}}$. In this case, $\tau(v) = \lambda v$ and $\tau(\sigma v) = \lambda^2 \sigma v$. Since $\lambda \neq \lambda^2$, the eigenvectors v and σv are linearly independent, and their span forms an invariant subspace. By irreducibility, this span must equal W, and we conclude that $W \simeq V$.

□

Given a representation $W \simeq U^{\oplus a} \oplus U'^{\oplus b} \oplus V^{\oplus c}$, how do we find a, b, c?

Answer: Look at the eigenvalues of τ. The 1-eigenspace of τ corresponds to $U^{\oplus a} \oplus U'^{\oplus b}$, so $a + b = \dim(\mathrm{Ker}(\tau - 1))$. The eigenvalue $e^{\pm \frac{2\pi i}{3}}$ corresponds to the V-eigenspace, so the multiplicity of these eigenvalues gives c. Similarly, σ acts by $+1$ on U, -1 on U', and by a matrix resembling the standard action on V. From this, we deduce a, b, c.

Example 5.16. *Consider the standard representation V of S_3 and $V^{\otimes 2} = V \otimes V$. How does $V^{\otimes 2}$ decompose into irreducibles?*

Start with a basis e_1, e_2 of V, where $\tau e_1 = \lambda e_1$, $\tau e_2 = \lambda^2 e_2$, $\sigma e_1 = e_2$, and $\sigma e_2 = e_1$, with $\lambda = e^{\frac{2\pi i}{3}}$. Then $V \otimes V$ has a basis $e_1 \otimes e_1, e_1 \otimes e_2, e_2 \otimes e_1, e_2 \otimes e_2$. These are eigenvectors of τ with eigenvalues $\lambda^2, 1, 1, \lambda$. On the 1-eigenspace

span$(e_1 \otimes e_2, e_2 \otimes e_1)$*, σ swaps the two vectors, so $e_1 \otimes e_2 \pm e_2 \otimes e_1$ is an eigenvector of σ with eigenvalue ± 1. Hence, $V \otimes V \simeq U \oplus U' \oplus V$.*

Similarly, $Sym^2(V)$ has a basis e_1^2, e_1e_2, e_2^2, and $Sym^2(V) \simeq U \otimes V$. The action of τ on this space is given by eigenvalues $\lambda^2, 1, \lambda$, while $\wedge^2 V \simeq U'$, corresponding to the determinant versus the sign.

Next, we will discuss symmetric polynomials and introduce characters as a tool to study representations.

5.3 Symmetric Polynomials and Characters

This concept generalizes to more complicated groups. We'll see that eigenvalues play a crucial role in classifying representations. However, we need a systematic way to organize this information.

Now, we digress to discuss symmetric polynomials, which serves as motivation for the study of characters.

Observe that an efficient way to store information about n complex numbers, unordered and possibly with repetitions, is to specify the coefficients of the polynomial whose roots are these numbers. That is, the polynomial is of the form $\prod_{i=1}^n (x - \lambda_i)$. The coefficients of this polynomial are **symmetric polynomials** in the variables $\lambda_1, \dots, \lambda_n$.

The symmetric group S_n acts on the space of polynomials $\mathbb{C}[z_1, \dots, z_n]$ by permuting the variables.

Definition 5.17. *A **symmetric polynomial** is a polynomial $f \in \mathbb{C}[z_1, \dots, z_n]$ that is a fixed point of the S_n-action, i.e., $\sigma(f) = f$ for all $\sigma \in S_n$.*

Remark 5.18. *Equality of polynomials means, as usual, equality of coefficients. Over a finite field, this is a stronger condition than equality as functions on k^n, but there is no distinction over $\mathbb{C}$.*

Definition 5.19. *The **elementary symmetric polynomials** are defined as follows:*

$$\sigma_1(z_1, \dots, z_n) = \sum_{i=1}^n z_i,$$

$$\sigma_2(z_1, \dots, z_n) = \sum_{1 \le i < j \le n} z_i z_j,$$

$$\vdots$$

$$\sigma_k(z_1, \dots, z_n) = \sum_{1 \le i_1 \le \cdots \le i_k \le n} z_{i_1} \dots z_{i_k},$$

$$\sigma_n = \prod_{i=1}^n z_i.$$

We can check that the coefficient of x^{n-k} in the polynomial $\prod_{i=1}^{n}(x - z_i)$ is, up to a sign factor of $(-1)^k$, the elementary symmetric polynomial $\sigma_k(z_1, \ldots, z_n)$.

Thus, by the fundamental theorem of algebra, there is a bijection

$$\{\text{unordered } n-\text{tuples of complex numbers, with repetitions allowed}\} \stackrel{\sim}{\Longleftrightarrow} \mathbb{C}^n \text{ ordered tuples}$$

given by

$$[z_1, \ldots, z_n] \mapsto (\sigma_1(z_i), \ldots, \sigma_n(z_i)),$$

and

$$(\sigma_1, \ldots, \sigma_n) \mapsto [\text{the roots of } x^n - \sigma_1 x^{n-1} + \cdots + (-1)^n \sigma_n].$$

In other words, $[z_1, \ldots, z_n] \Longleftrightarrow$ the coefficients of the polynomial $\prod(x - z_i)$.

Theorem 5.20. *The subring of symmetric polynomials in $\mathbb{C}[z_1, \ldots, z_n]$, i.e., $\mathbb{C}[z_1, \ldots, z_n]^{S_n}$, is isomorphic to the polynomial algebra in the elementary symmetric polynomials.*

We won't prove this here, but let's see why it works in the case $n = 2$:

The vector space of symmetric polynomials has the following basis:

$$\begin{aligned}
1 &= 1, \\
z_1 + z_2 &= \sigma_1, \\
z_1^2 + z_2^2 &= (z_1 + z_2)^2 - 2z_1 z_2 = \sigma_1^2 - \sigma_2, \\
z_1 z_2 &= \sigma_2, \\
z_1^3 + z_2^3 &= \sigma_1^3 - 3z_1^2 z_2 - 3z_1 z_2^2 = \sigma_1^3 - 3\sigma_1\sigma_2, \\
z_1^2 z_2 + z_1 z_2^2 &= \sigma_1\sigma_2.
\end{aligned}$$

Observe that any symmetric polynomial in 2 variables can be written as

$$\begin{aligned}
p(z_1, z_2) &= \sum a_k (z_1^k + z_2^k) + z_1 z_2 q(z_1, z_2) \\
&= \sum a_k (z_1 + z_2)^k + z_1 z_2 q'(z_1, z_2) \\
&= \sum a_k \sigma_1^k + \sigma_2 q'.
\end{aligned}$$

We can continue by induction on the degree.

Remark 5.21. *The theorem can be understood in terms of the representation theory of S_n! Specifically, the space of homogeneous degree 1 polynomials is $W_1 = span(z_1, \ldots, z_n) \simeq \mathbb{C}^n$, on which S_n acts by the permutation representation, which decomposes as $V \oplus U$, with the invariant part $W_1^{S_n} \simeq U$ being the trivial summand. Similarly, homogeneous degree d polynomials correspond*

to $W_d = Sym^d(W_\ell)$, and the invariant part $W_d^{S_n}$ is the trivial summand in the decomposition of W_d into irreducibles. (Unfortunately, we haven't studied the representations of S_n in enough depth to carry through with a proof along these lines.)

Another family of symmetric polynomials are the power sums:

$$\tau_k(z_1, \dots, z_n) = \sum_{i=1}^{n} z_i^k,$$

where $\tau_1 = \sigma_1, \tau_2 = \sigma_1^2 - 2\sigma_2, \dots$.

These make sense for all k, but in fact, $\tau_1, \dots, \tau_n$ are sufficient:

Theorem 5.22.

$$\mathbb{C}[z_1, \dots, z_n]^{S_n} \cong \mathbb{C}[\tau_1, \dots, \tau_n].$$

In particular, specifying an unordered tuple $\{z_1, \dots, z_n\}$ is equivalent to specifying $\sum z_i, \sum z_i^2, \dots, \sum z_i^n$.

Now, let's return to representation theory to see why this matters. We've seen that to understand a representation V of a group G, we should examine the eigenvalues of the action $g : V \to V$ for each $g \in G$. However, this provides a lot of information. We've just established that to specify the eigenvalues λ_i of $g : V \to V$, it is enough to specify the power sums $\sum \lambda_i^k$. In fact, $\sum \lambda_i^k = \operatorname{tr}(g^k)$. Thus, it suffices to describe just the sum of the eigenvalues, $\sum \lambda_i = \operatorname{tr}(g)$, for each $g \in G$, since G is a group, and the trace of g^k is also part of this.

Definition 5.23. *The **character** χ_V of a representation V is the function $\chi_V : G \to \mathbb{C}$ defined by $\chi_V(g) = tr(g)$.*

Remark 5.24. *For a 1-dimensional representation of G, i.e., a homomorphism $G \to \mathbb{C}^*$, the character is just the same as the homomorphism. For a higher-dimensional representation, however, we have $\chi(g_1 g_2) \neq \chi(g_1)\chi(g_2)$.*

However, since trace is conjugation-invariant, i.e., $\operatorname{tr}(ghg^{-1}) = \operatorname{tr}(h)$, the character $\chi_V(g)$ only depends on the conjugacy class of g.

Definition 5.25. *A **class function** $f : G \to \mathbb{C}$ is a function that is invariant under conjugation, i.e., $f(ghg^{-1}) = f(h)$.*

Example 5.26. *Given representations V and W:*

- $\chi_{V \oplus W}(g) = \chi_V(g) + \chi_W(g)$ *(eigenvalues of the block matrix $\begin{pmatrix} \varphi & 0 \\ 0 & \varphi \end{pmatrix}$),*
- $\chi_{V \otimes W}(g) = \chi_V(g)\chi_W(g)$ *(eigenvalues of $\varphi \otimes \psi : v_i \otimes w_j \mapsto \lambda_i \lambda'_j v_i \otimes w_j$),*
- $\chi_{V^*}(g) = \overline{\chi_V(g)}$ *since g acts by $t(g^{-1})$, and the eigenvalues are roots of unity, so $\lambda_i^{-1} = \overline{\lambda_i} \implies \sum \lambda_i^{-1} = \sum \overline{\lambda_i}$,*
- $\chi_{\wedge^2 V}(g) = \sum_{i<j} \lambda_i \lambda_j = \frac{1}{2}\left((\sum \lambda_i)^2 - \sum \lambda_i^2\right) = \frac{1}{2}(\chi_V(g)^2 - \chi_V(g^2))$.

Example 5.27. *If G acts on a finite set S, then there is an associated permutation representation V of dimension $|S|$, with basis $(e_s)_{s\in S}$, where G acts by permutation matrices $g{\cdot}e_s = e_{g\cdot s}$. In this case, $\chi_V(g) = tr(g) = \#\{s \in S \mid g{\cdot}s = s\}$, since the 1's on the diagonal of the matrix correspond to fixed points of g, and the 0's correspond to non-fixed points.*

The **character table** of a group is the table listing, for each irreducible representation of G, the values of its character on each conjugacy class of G.

Example 5.28. *For $G = S_3$, the character table is:*

S_3	e	$(1\,2)$	$(1\,2\,3)$
U	1	1	1
U'	1	-1	1
V	2	0	-1

where e, $(1\,2)$, and $(1\,2\,3)$ represent the conjugacy classes, and U, U', and V represent the irreducible representations.

For the left column, $\chi_V(e) = \mathrm{tr}(\mathrm{id}) = \dim(V)$, and for the bottom row, the eigenvalues are ± 1 for $(1\,2)$, $e^{\frac{2\pi i}{3}}$ for $(1\,2\,3)$, or $U \oplus V =$ permutation representation, which takes values $(3, 1, 0)$; subtracting $\chi_U = (1, 1, 1)$ gives the result.

We now have a faster way to decompose $V \otimes V$ into its irreducibles: $\chi_{V\otimes V}(g) = \chi_V(g)^2$, so $\chi_{V\otimes V}$ takes values $(4, 0, 1)$. Since χ_U, $\chi_{U'}$, and χ_V are linearly independent, $\chi_{V\otimes V} = \chi_U + \chi_{U'} + \chi_V$, implying $V \otimes V \simeq U \oplus U' \oplus V$. This method is faster than counting eigenvalues, as we did previously!

Now for some magic with characters: If V is a representation of G, the invariant part is

$$V^G = \{v \in V \mid gv = v\, \forall g \in G\}.$$

Proposition 5.29. *The map*

$$\varphi = \frac{1}{|G|} \sum_{g\in G} g : V \to V$$

is a projection onto $V^G \subset V$, satisfying $Im(\varphi) = V^G$ and $\varphi|_{V^G} = id$.

Thus,

$$\dim(V^G) = \mathrm{tr}(\varphi) = \frac{1}{|G|} \sum_{g\in G} \chi_V(g).$$

If V, W are representations of G, then

$$\mathrm{Hom}_G(V, W) = \mathrm{Hom}(V, W)^G = (V^* \otimes W)^G,$$

which implies

$$\dim(\mathrm{Hom}_G(V,W)) = \frac{1}{|G|}\sum_{g\in G}\chi_{V^*\otimes W}(g) = \frac{1}{|G|}\sum_{g\in G}\overline{\chi_V(g)}\chi_W(g).$$

If V and W are irreducible, then by Schur's lemma:

$$\dim(\mathrm{Hom}_G(V,W)) = \begin{cases} 1 & \text{if } V \simeq W, \\ 0 & \text{otherwise.} \end{cases}$$

Definition 5.30. *Define a Hermitian inner product on the space of class functions $G \to \mathbb{C}$ by*

$$H(\alpha,\beta) = \frac{1}{|G|}\sum_{g\in G}\overline{\alpha(g)}\beta(g).$$

For characters of representations, by the above,

$$\dim(\mathrm{Hom}_G(V,W)) = H(\chi_V,\chi_W).$$

This leads to the following theorem:

Theorem 5.31. *The characters of irreducible representations of G are orthonormal with respect to H.*

As a consequence, the characters of irreducible representations are linearly independent class functions.

Corollary 5.32. *The number of irreducible representations of G is at most the number of conjugacy classes of G.*

Remark 5.33. *We will see later that these numbers are in fact equal.*

Corollary 5.34. *Every representation of G is completely determined by its character. Denoting the irreducible representations by $V_1,\dots,V_k$, any representation W can be expressed as*

$$W \simeq \bigoplus V_i^{\oplus a_i},$$

where

$$a_i = \dim(Hom_G(V_i,W)) = H(\chi_{V_i},\chi_W).$$

Corollary 5.35. *For any representation $W = \bigoplus V_i^{\oplus a_i}$,*

$$H(\chi_W,\chi_W) = \sum a_i^2,$$

and W is irreducible if and only if $H(\chi_W,\chi_W) = 1$.

This is particularly useful because, given a representation W, it provides information about the number of irreducible summands in W. For example:

$$H(\chi_W, \chi_W) = \begin{cases} 1 & \iff W \text{ is irreducible,} \\ 2 & \iff W \text{ is a direct sum of 2 different irreducible representations,} \\ 4 & \iff W \text{ is either 4 different irreducible representations or twice the same.} \end{cases}$$

We now apply this to the regular representation R, which is the vector space with basis $\{e_g\}_{g\in G}$, where G acts by permuting the basis vectors by left multiplication: $g \cdot e_h = e_{gh}$.

Let $V_1, \ldots, V_k$ be the irreducible representations of G, and write

$$R = \bigoplus V_i^{\oplus a_i}.$$

What are the a_i?

Since G acts by permutation matrices,

$$\chi_R(g) = \operatorname{tr}(g) = \#\{h \in G \mid g \cdot e_h = e_h\}.$$

Unless $g = e$, there are no fixed points, which implies:

$$\chi_R(g) = \begin{cases} |G| & \text{if } g = e, \\ 0 & \text{if } g \neq e. \end{cases}$$

Thus,

$$H(\chi_R, \chi_{V_i}) = \frac{1}{|G|} \sum_{g\in G} \overline{\chi_R(g)} \chi_{V_i}(g) = \chi_{V_i}(e) = \operatorname{tr}(\mathrm{id}_{V_i}) = \dim(V_i).$$

Hence, each V_i appears $a_i = \dim(V_i)$ times in the regular representation R. The third corollary then implies:

$$H(\chi_R, \chi_R) = \frac{1}{|G|} \sum_{g\in G} |\chi_R(g)|^2 = \frac{1}{|G|} |\chi_R(e)|^2 = |G| = \sum a_i^2 = \sum (\dim(V_i))^2.$$

Corollary 5.36. *The irreducible representations $V_1, \ldots, V_k$ of G satisfy*

$$\sum (\dim(V_i))^2 = |G|.$$

At this point, we have significant information about the irreducible representations of G and their characters.

5.4 S_4

Let $G = S_4$. The conjugacy classes of S_4 are:

$\{e\}$, size 1, transpositions, size 6, 3-cycles, size 8, 4-cycles, size 6, pairs of transpositions, size

We know three irreducible representations: U, U', and V. The character table is as follows:

S_4	e	$(1\,2)$	$(1\,2\,3)$	$(1\,2\,3\,4)$	$(1\,2)(3\,4)$
U	1	1	1	1	1
U'	1	-1	1	-1	1
V	3	1	0	-1	-1

For the first row, g acts as the identity, so the trace is tr = 1. For the second row, we have $\mathrm{tr}((-1)^\sigma) = (-1)^\sigma$, meaning the trace is either 1 or -1, depending on the permutation. For the last row, the direct sum $U \oplus V$ is the permutation representation, $\mathbb{C}^4$, where $\chi_{U\oplus V}(\sigma) = \mathrm{tr}(\sigma)$ represents the number of fixed points of σ, i.e., the number of i such that $\sigma(i) = i$. This implies $\chi_V(\sigma) =$ # fixed points $- 1$.

Quick check: these characters are indeed orthonormal! However, we have $\sum \dim^2 = 1^2 + 1^2 + 3^2 = 11$, which is less than 24, so there must be other irreducible representations. In fact, by Corollary 1, we know there are at most two missing representations (since the number of irreducible representations is less than or equal to the number of conjugacy classes, which is 5). Since 13 is not a perfect square, we conclude that exactly two irreducible representations are missing, with dimensions 2 and 3.

How do we build the missing entries? We begin by looking at tensor products of known irreducible representations. First, recall that the tensor product of an irreducible representation with a 1-dimensional representation is still irreducible, as it leaves the same invariant subspaces. So, we consider the tensor product $V' = V \otimes U'$, which represents the twice-standard representation by $(-1)^\sigma$. The character $\chi_{V'}$ is given by:

$$\chi_{V'} = \chi_V \cdot \chi_{U'} = (3, -1, 0, 1, -1),$$

which is irreducible, as $H(\chi_{V'}, \chi_{V'}) = 1$, and different from V.

Thus, we have found one of the missing irreducible representations, V'.

Now, we find the last 2-dimensional irreducible representation, W. Since $W \otimes U'$ is also 2-dimensional and irreducible, we deduce that $W \otimes U' \simeq W$. This implies $\chi_W = \chi_W \cdot \chi_{U'}$, so $\chi_W = 0$ for the odd conjugacy classes $(1\,2)$ and $(1\,2\,3\,4)$. The orthogonality relations help us determine the rest of χ_W without explicitly

constructing it. The character table now becomes:

S_4	e	$(1\,2)$	$(1\,2\,3)$	$(1\,2\,3\,4)$	$(1\,2)(34)$
U	1	1	1	1	1
U'	1	-1	1	-1	1
V	3	1	0	-1	-1
V'	3	-1	0	1	-1
W	2	0	$a=-1$	0	$b=2$

To find a and b, we use the orthogonality relations:

$$H(\chi_V,\chi_W)=\frac{1}{24}(2{+}8a{+}3b)=0,\quad H(\chi_V,\chi_W)=\frac{1}{24}(6{-}3b)=0\quad\Longrightarrow\ (a,b)=(-1,2).$$

Note that $\chi_W((1\,2)(34))=2$, which means the eigenvalues of W are 1 and 1 (the roots of unity summing to 2).

This gives us a clue about W: the normal subgroup $H=\{\mathrm{id}\}\cup\{(ij)(kl)\}\simeq\mathbb{Z}/2\times\mathbb{Z}/2$ lies in the kernel of the map $S_4\xrightarrow{\rho}\mathrm{GL}(W)$. In other words, ρ factors through the quotient $S_4/H\simeq S_3$, where S_4 acts on the set of splittings of $\{1,2,3,4\}$ into two pairs (there are three such splittings). Under this quotient:

- Transpositions and 4-cycles map to transpositions.
- 3-cycles map to 3-cycles.

Thus, the character χ_W on the quotient S_3 becomes:

$$\begin{cases}\mathrm{id} & \mapsto 2,\\ \text{transposition} & \mapsto 0,\\ \text{3-cycle} & \mapsto -1.\end{cases}$$

This is the standard representation of S_3, "pulled back" to S_4 by the map $S_4\twoheadrightarrow S_3$.

Alternatively, we can look at the tensor product $V\otimes V$. The character of this product is $\chi_{V\otimes V}=\chi_V^2=(9,1,0,1,1)$. Using the orthogonality relations:

$$H(\chi_U,\chi_{V\otimes V})=1,\quad H(\chi_{U'},\chi_{V\otimes V})=0,\quad H(\chi_V,\chi_{V\otimes V})=\frac{1}{24}(27{+}6{-}6{-}3)=1,$$

$$H(\chi_{V'},\chi_{V\otimes V})=\frac{1}{24}(27-6+6-3)=1,$$

we find that $V\otimes V$ contains $U\oplus V\oplus V'$ (dimension 7), leaving one copy of the missing irreducible W. Therefore:

$$V\otimes V=U\oplus V\oplus V'\oplus W.$$

We can find χ_W by subtracting the other characters from $\chi_{V\otimes V}$.

5.5 A_4

The alternating group A_4 has 4 conjugacy classes:

$\{e\}$ (1 element), $(1\,2\,3)$ (4 elements), $(1\,3\,2))$ (4 elements), $(1\,2)(34)$ (3 elements).

We can approach the classification of irreducible representations of A_4 by first restricting the irreducible representations of S_4 to A_4. Some of these become isomorphic (for instance, the alternating representation U' has elements of A_4 acting by $(-1)^\sigma = 1$, so it is isomorphic to the trivial representation), while others may become reducible. This approach is feasible but challenging, particularly due to the role of the representation W.

Alternatively, we can proceed directly. We know that A_4 has at most 4 irreducible representations, and the sum of the squares of their dimensions must equal 12, i.e.,

$$\sum \dim^2 = 12.$$

This includes the trivial representation of dimension 1, so the only possible dimension configuration is:

$$12 = 3^2 + 1^2 + 1^2 + 1^2.$$

The three 1-dimensional representations correspond to $\mathrm{Hom}(A_4, \mathbb{C}^*)$, which contains the trivial representation and two other elements.

Now, observe that $H = \{\mathrm{id}\} \cup \{(ij)(kl)\}$ is a normal subgroup of A_4, and $A_4/H \simeq \mathbb{Z}/3$. Thus, we can conclude that:

$$\mathrm{Hom}(A_4, \mathbb{C}^*) \simeq \widehat{\mathbb{Z}/3} = \left\{ m \mapsto e^{\frac{2\pi i m k}{3}} \right\}.$$

Specifically, let $\lambda = e^{\frac{2\pi i}{3}}$, and the rank-1 representations are as follows:

A_4	e	$(1\,2\,3)$	$(1\,3\,2))$	$(1\,2)(34)$
U	1	1	1	1
U'	1	λ	λ^2	1
U''	1	λ^2	λ	1
V	3	0	0	-1

Remark 5.37. *Note that the restriction of the standard representation W of S_4 to A_4 decomposes as $W|_{A_4} \simeq U' \oplus U''$. Elements of H (i.e., $(ij)(kl)$) act as the identity, and this is the restriction of the standard representation of S_4.*

In the previous section, we stated (but did not prove) that the characters of irreducible representations form an orthonormal basis (with respect to the inner product for class functions) of the space of class functions $G \to \mathbb{C}$.

The proof follows from a general averaging/projection formula. As we saw earlier, the operator $\varphi = \frac{1}{|G|} \sum_{g \in G} g$ is a projection onto the invariant subspace V^G (the trivial summand in V).

Proposition 5.38. *Given any class function $\alpha : G \to \mathbb{C}$ and any representation V of G, let*

$$\varphi_{\alpha,V} = \frac{1}{|G|}\sum_{g\in G}\alpha(g)g : V \to V.$$

Then $\varphi_{\alpha,V}$ is G-linear (equivariant).

Proof. We compute:

$$\varphi_{\alpha,V}(hv) = \frac{1}{|G|}\sum_{g\in G}\alpha(g)ghv \quad = \frac{1}{|G|}\sum_{g'\in G}\alpha(hg'h^{-1})(hg'h^{-1})hv = \frac{1}{|G|}\sum_{g'\in G}\alpha(g')hg'v \quad = h\left(\frac{1}{|G|}\sum_{g'\in G}\alpha(g')g'v\right.$$

Thus, $\varphi_{\alpha,V}$ is G-linear. □

This leads to the following theorem:

Theorem 5.39. *The characters of the irreducible representations of G form an orthonormal basis (with respect to the inner product for class functions) of the space of class functions $G \to \mathbb{C}$, and the number of irreducible representations equals the number of conjugacy classes.*

Proof. To show that the characters $\chi_1, \ldots, \chi_m$ of the irreducible representations span the space of class functions, it suffices to prove that $H(\overline{\alpha}, \chi_i) = 0$ for all i implies that $\alpha = 0$. Given any class function α and an irreducible representation V, the operator $\varphi_{\alpha,V}$ is defined as above. By Schur's Lemma, $\varphi_{\alpha,V} = \lambda \mathrm{id}_V$, where $\lambda = \frac{1}{n}\mathrm{tr}(\varphi_{\alpha,V})$ and $n = \dim(V)$. Thus:

$$\lambda = \frac{1}{n}\mathrm{tr}(\varphi_{\alpha,V}) = \frac{1}{n}\frac{1}{|G|}\sum_{g\in G}\alpha(g)\mathrm{tr}(g) = \frac{1}{n}\frac{1}{|G|}\sum\alpha(g)\chi_V(g) = \frac{1}{n}H(\overline{\alpha}, \chi_V).$$

So if $H(\overline{\alpha}, \chi_{V_i}) = 0$ for all irreducible representations V_i, then $\varphi_{\alpha,V_i} = 0$ for all V_i. By considering direct sums, we have $\varphi_{\alpha,V} = 0$ for all representations of G, particularly for the regular representation R of G (the permutation representation for left multiplication on G).

For the regular representation, we compute:

$$\varphi_{\alpha,R}(e_1) = \frac{1}{|G|}\sum_{g\in G}\alpha(g)e_g = 0.$$

Since the e_g are linearly independent, this implies $\alpha(g) = 0$ for all $g \in G$, so $\alpha = 0$. □

Along the way, we find the following result. For irreducible representations V_i and V_j, consider $\varphi_{\alpha,V}$ for $\alpha = \overline{\chi_{V_i}}$. Then $\varphi_{\alpha,V_j} = \lambda \mathrm{id}_{V_j}$, where:

$$\lambda = \frac{1}{\dim(V_j)}\mathrm{tr}(\varphi_{\alpha,V_j}) = \frac{1}{\dim V_j}H(\chi_{V_i}, \chi_{V_j}) = \begin{cases}\frac{1}{\dim V_j}, & \text{if } i = j,\\ 0, & \text{if } i \neq j.\end{cases}$$

This leads to the following proposition:

Proposition 5.40. *If V is any representation of G, and $V = \bigoplus V_i^{\oplus a_i}$ is its decomposition into irreducibles, then:*

$$\varphi_{V_i} = \frac{\dim V_i}{|G|} \sum_{g \in G} \overline{\chi_{V_i}(g)} g : V \to V$$

is the projection onto the summand $V_i^{\oplus a_i}$ (i.e., it is the identity on that summand and zero on the others).

In the case of the trivial representation, this formula reduces to our previous projection formulas for V^G.

5.6 The Representation Ring of G

Fix a group G and consider the set of representations of G up to isomorphism. There are two operations, $\oplus$ and $\otimes$, which are commutative, associative, and distributive:

$$(U \oplus V) \otimes W = (U \otimes W) \oplus (V \otimes W).$$

So, is this a ring? ... Almost! We are missing additive inverses — let's just add them! Define

$$\hat{R} = \left\{ \sum_{\text{finite}} a_i [V_i] \mid a_i \in \mathbb{Z}, V_i \text{ are representations of } G \right\},$$

the set of formal linear combinations with integer coefficients of representations of G. Consider the additive subgroup generated by all sums $[V]+[W] = [V \oplus W]$. Let $R(G)$ be the quotient of $\hat{R}$ by this subgroup. In $R(G)$, we have $[V] + [W] = [V \oplus W]$, but we can also subtract representations.

Definition 5.41. *The pair $(R(G), \oplus, \otimes)$ is called the **representation ring** of G, where we extend these operations to formal sums and differences of representations by linearity.*

As a set, $R(G)$ consists of elements of the form

$$R(G) = \left\{ \sum_{i=1}^{k} a_i V_i \mid a_i \in \mathbb{Z} \right\},$$

where V_i are the irreducible representations of G (by complete reducibility and the uniqueness of decomposition into irreducible representations). In other words, $(R(G), +)$ is a free abelian group, isomorphic to $\mathbb{Z}^k$, where k is the number of irreducible representations of G.

General elements of $R(G)$, i.e., the coefficients $a_i \in \mathbb{Z}$, are called *virtual representations*. Actual representations, i.e., elements such that $a_i \geq 0$ for all i, form a cone inside $R(G)$ (i.e., a subset that is closed under addition).

Next, the character map $V \mapsto \chi_V$ can be extended by linearity to a map $R(G) \to \mathbb{C}_{\text{class}}(G)$, where $\mathbb{C}_{\text{class}}(G)$ denotes the space of class functions on G. This map is a ring homomorphism, since

$$\chi_{U\oplus V} = \chi_U + \chi_V \quad \text{and} \quad \chi_{U\otimes V} = \chi_U \chi_V.$$

The image of this map consists of the *virtual characters*, which are formal linear combinations of the irreducible characters:

$$\left\{ \sum a_i \chi_{V_i} \mid a_i \in \mathbb{Z} \right\}.$$

If we pass to complex linear combinations instead of integer coefficients, our results about irreducible characters forming a basis imply that:

$$R(G) \otimes_{\mathbb{Z}} \mathbb{C} \simeq \mathbb{C}_{\text{class}}(G),$$

where the map is given by

$$\sum_{i=1}^{k} a_i [V_i] \mapsto \chi_{\sum a_i V_i} = \sum a_i \chi_{V_i}.$$

This is an isomorphism, since the tensor product of free $\mathbb{Z}$-modules behaves similarly to the tensor product of vector spaces.

There are theorems of Artin and Brauer that describe the lattice of virtual characters

$$\wedge = \left\{ \sum a_i \chi_{V_i} \mid a_i \in \mathbb{Z} \right\}$$

inside $\mathbb{C}_{\text{class}}(G)$. We will explore these results later.

5.7 S_5

Now, let's explore the representations of S_5 and A_5 to gain further practice with characters and to motivate the discussion of restriction and induction of representations (representations of $G \iff$ representations of subgroups of G).

One can begin constructing the character table of S_5 in the usual manner: start with known representations. We know that $V \oplus U \simeq$ the permutation representation $\mathbb{C}^5$, so $\chi_{V\oplus U}(\sigma) = \#\{i \mid \sigma(i) = i\}$, and similarly for χ_V, $\chi_{U\oplus V}$, with the shift by -1.

S_5	e	$(1\,2)$	$(1\,2\,3)$	$(1\,2\,3\,4)$	(12345)	$(1\,2)(34)$	$(1\,2\,3)(45)$
U	1	1	1	1	1	1	1
U'	1	-1	1	-1	1	1	-1
V	4	2	1	0	-1	0	-1
$V' = V \otimes U'$	4	-2	1	0	-1	0	1

Next, we need to find more irreducible representations. Since $|S_5| = 120 = \sum \dim^2$, we are still missing 3 irreducibles with $\sum \dim^2 = 86$. The most effective way to find them is to continue building tensor products — namely, consider $V \otimes V$ (dimension 16), or more specifically, its two parts: $\mathrm{Sym}^2(V)$ (dimension 10) and $\wedge^2 V$ (dimension 6).

Observe that if $g : V \to V$ has eigenvalues λ_i ($gv_i = \lambda_i v_i$, for $1 \le i \le r$), then the corresponding maps on $\mathrm{Sym}^2(V)$ have eigenvalues $\lambda_i \lambda_j$ for $1 \le i \le j \le r$, because (v_i) forms a basis for $V \implies (v_i v_j)$ forms a basis for $\mathrm{Sym}^2(V)$. Similarly, $\wedge^2(V)$ has eigenvalues $\lambda_i \lambda_j$ for $1 \le i < j \le r$, since (v_i) forms a basis for $V \implies (v_i \wedge v_j)$ forms a basis for $\wedge^2(V)$.

Now, we have the following identities:

$$\sum_{i \le j} \lambda_i \lambda_j = \frac{1}{2}\left(\left(\sum \lambda_i\right)^2 - \sum \lambda_i^2\right),$$
$$\implies \chi_{\mathrm{Sym}^2(V)}(g) = \frac{1}{2}\left(\chi_V(g)^2 - \chi_V(g^2)\right),$$

and

$$\sum_{i < j} \lambda_i \lambda_j = \frac{1}{2}\left(\left(\sum \lambda_i\right)^2 - \sum \lambda_i^2\right),$$
$$\implies \chi_{\wedge^2(V)}(g) = \frac{1}{2}\left(\chi_V(g)^2 - \chi_V(g^2)\right).$$

These formulas allow us to calculate $\chi_{\mathrm{Sym}^2(V)}$ and $\chi_{\wedge^2(V)}$ for the standard representation of S_5.

S_5	e	$(1\,2)$	$(1\,2\,3)$	$(1\,2\,3\,4)$	(12345)	$(1\,2)(34)$	$(1\,2\,3)(45)$
V	4	2	1	0	-1	0	-1
$\wedge^2(V)$	6	0	0	0	1	-2	0
$\mathrm{Sym}^2(V)$	10	4	1	0	0	2	1

Observe that

$$H(\chi_{\wedge^2(V)}, \chi_{\wedge^2(V)}) = \frac{1}{120}\left(6^2 + 24 + 15 \cdot 2^2\right) = 1,$$

so $\wedge^2(V)$ is irreducible! On the other hand,

$$H(\chi_{\mathrm{Sym}^2(V)}, \chi_{\mathrm{Sym}^2(V)}) = \frac{1}{120}\left(10^2 + 10 \cdot 4^2 + 20 + 15 \cdot 2^2 + 20\right) = 3,$$

so $\mathrm{Sym}^2(V)$ decomposes into 3 irreducible summands.

Additionally,

$$H(\chi_V, \chi_{\mathrm{Sym}^2(V)}) = \frac{1}{120}\left(10^2 + 10 \cdot 4 + 20 + 15 \cdot 2 + 20\right) = 1,$$

so $\mathrm{Sym}^2(V)$ contains one copy of U. Similar calculations show that $\mathrm{Sym}^2(V)$ also contains V with multiplicity 1, but not U' or V'.

Therefore, $\mathrm{Sym}^2(V) = U \oplus V \oplus W$, where W is some irreducible 5-dimensional representation. Subtracting the known representations, we find χ_W, and by considering $W' = W \otimes U'$, we complete the list of irreducible representations.

Thus, we have the table:

S_5	e	(12)	(123)	(1234)	(12345)	$(12)(34)$	$(123)(45)$
U	1	1	1	1	1	1	1
U'	1	-1	1	-1	1	1	-1
V	4	2	1	0	-1	0	-1
$V' = V \otimes U'$	4	-2	1	0	-1	0	1
$\wedge^2(V)$	6	0	0	0	1	-2	0
W	5	1	-1	-1	0	1	1
$W' = W \otimes U'$	5	-1	-1	1	0	1	-1

Remark 5.42. *The standard representation V and its exterior powers, $\wedge^2(V)$, $\wedge^3(V) \simeq V'$, and $\wedge^4(V) \simeq U'$ are all irreducible! This is, in fact, a general property: for all $0 \leq k \leq n-1$, the exterior powers $\wedge^k(V)$ of the standard representation of S_n are irreducible.*

5.8 A_5

Now let's move on to A_5. We begin by restricting the irreducible representations of S_5 to A_5 and examining which ones remain irreducible or decompose. Naturally, different irreducible representations of S_5 can become isomorphic after restriction. Specifically, elements of A_5 act trivially on U', so U' becomes trivial, and the restrictions of V and $V' = V \otimes U'$ become isomorphic. Similarly, W also decomposes. The character table for S_5 gives, after restriction, the following:

A_5	e	$(1\,2\,3)$	(12345)	(12354)	$(1\,2)(34)$
U	1	1	1	1	1
V	4	1	-1	-1	0
W	5	-1	0	0	1
$\wedge^2(V)$	6	0	1	1	-2

By calculating $H(\chi, \chi)$, we find that U, V, and W are irreducible, while $H(\chi_{\mathrm{Sym}^2(V)}, \chi_{\mathrm{Sym}^2(V)}) = 2$, so $\wedge^2(V)$ decomposes into the direct sum of two distinct irreducible representations. Additionally, $\wedge^2(V)$ does not contain U, V, or W, so we conclude that

$\wedge^2(V) = Y \oplus Z$, where Y and Z are the last two irreducible representations of A_5.

From the fact that the sum of the squared dimensions equals the order of A_5, $\sum \dim^2 = |A_5| = 60$, we deduce that $\dim(Y) = \dim(Z) = 3$. To determine χ_Y and χ_Z, we use orthogonality. Since $\chi_Y + \chi_Z = \chi_{\wedge^2(V)}$, we find that $\chi_Y - \chi_Z$ lies in the orthogonal complement of the span of $\chi_U, \chi_V, \chi_W, \chi_{\wedge^2(V)}$. Hence, $\chi_Y - \chi_Z = (0, 0, a, -a, 0)$, where $H(\chi_Y - \chi_Z, \chi_Y - \chi_Z) = 2$. This implies:

$$24a^2 = 120 \quad \Longrightarrow \quad a = \pm\sqrt{5}.$$

Thus, the characters of Y and Z are:

A_5	e	(123)	(12345)	(12354)	$(12)(34)$
U	1	1	1	1	1
V	4	1	-1	-1	0
W	5	-1	0	0	1
Y	3	0	$\frac{1+\sqrt{5}}{2}$	$\frac{1-\sqrt{5}}{2}$	-1
Z	3	0	$\frac{1-\sqrt{5}}{2}$	$\frac{1+\sqrt{5}}{2}$	-1

What are Y and Z? Recall that A_5 is the group of rotational symmetries of the icosahedron in $\mathbb{R}^3$. Hence, we have the inclusion $A_5 \hookrightarrow SO(3) \subset \mathrm{GL}(3, \mathbb{R}) \subset \mathrm{GL}(3, \mathbb{C})$. The representations Y and Z differ by an outer automorphism of A_5, specifically conjugation by a transposition inside S_5.

The fact that the character table contains irrational values implies that there does not exist a regular icosahedron (or dodecahedron) in $\mathbb{R}^3$ whose vertices all have rational coordinates. Otherwise, we would obtain a representation that factors through $\mathrm{GL}(3, \mathbb{Q})$, and we would have $\mathrm{tr}(g) \in \mathbb{Q}$ for all $g \in A_5$.

5.9 Induced Representations

We now present a more systematic approach: if G is a finite group and $H \subset G$ is a subgroup, then we define a restriction operation Res^G_H that maps representations of G to representations of H. This operation is, in fact, a functor $\mathrm{Rep}(G) \to \mathrm{Rep}(H)$, where objects are representations of G and H, and morphisms are homomorphisms (or equivalently, transpositions of representations). But what about the opposite direction?

Suppose V is a representation of G, and $W \subset V$ is invariant under H (but not necessarily under all of G). For $g \in G$, the subspace $gW \subset V$ depends only on the coset gH, and each gW is a representation of gHg^{-1}, with the diagram:

$$\begin{array}{ccc} H & \xrightarrow{\quad\rho\quad} & \mathrm{GL}(W) \\ {\scriptstyle c_g}\downarrow{\scriptstyle \simeq} & & \downarrow{\scriptstyle \text{conjugation by } g} \\ gHg^{-1} & \longrightarrow & \mathrm{GL}(gW) \end{array}$$

The simplest possible scenario is that

$$V = \bigoplus_{\sigma \in G/H} \sigma W$$

but in general, there is no reason for this to hold.

If it does hold, then the representation of G is completely determined by that of H. Indeed, choose representations $\sigma_1, \ldots, \sigma_k \in G$ for the cosets of H (each coset containing exactly one σ_i). Given $g \in G$, $g\sigma_i \in \sigma_j H$ for some j, so there exists $h \in H$ such that $g = \sigma_j h \sigma_i^{-1}$. Then g acts by mapping $\sigma_i W$ to $\sigma_j W$, with

$$g(\sigma_i w) = \sigma_j h(w).$$

Remark 5.43.

$$\dim(V) = |G/H| \cdot \dim(W).$$

Definition 5.44. *A representation V of G, with a subspace $W \subset V$ that is invariant under the subgroup $H \subset G$ (i.e., a subrepresentation of $Res_G^H(V)$), is said to be **induced** by $W \in Rep(H)$ if, as a vector space, $V = \bigoplus_{\sigma \in G/H} \sigma W$. We write $V = Ind_H^G W$.*

That is, by fixing one element in each coset $\sigma_1, \ldots, \sigma_k \in G$, we can write each $v \in V$ uniquely as

$$v = \sigma_1 w_1 + \cdots + \sigma_k w_k$$

where $w_1, \ldots, w_k \in W$.

Theorem 5.45. *Given a representation W of H, the induced representation $V = Ind_H^G W$ exists and is unique up to isomorphism of G-representations.*

Proof. **Uniqueness:** Given $V \in \mathrm{Rep}(G)$ and a subspace $W \subset V$ invariant under H such that $V = \bigoplus_{i=1}^k \sigma_i W$, the action of $g \in G$ maps $\sigma_i W$ to $\sigma_j W$, where j is such that $g\sigma_i \in \sigma_j H$, i.e., $h = \sigma_j^{-1} g \sigma_i \in H$. Thus, $g(\sigma_i w) = \sigma_j h w \in \sigma_j W$. This uniquely determines the action of g on V.

Existence: We build $V = \bigoplus_{i=1}^k \sigma_i W$, where the σ_i are formal symbols (i.e., the direct sum of $k = |G/H|$ copies of W), and we define the action of $g \in G$ as described above.

□

Example 5.46.

1. *The permutation representation associated with the left action of G on G/H is induced by the trivial representation of H. Indeed, V has a basis $\{e_\sigma\}_{\sigma \in G/H}$; the basis element e_H (for the coset H) is fixed by H,*

so $W = span(e_H)$ is invariant under H, and $gW = span(e_{gH})$, with $V = \bigoplus_{gH \in G/H} span(e_{gH}) = \bigoplus_{gH \in G/H} gW$.

2. *The regular representation of G is induced by the regular representation of H. Here, $W = span\{e_h : h \in H\} \subset V = span\{e_g : g \in G\}$.*

Proposition 5.47.

$$Ind_H^G(W \oplus W') = Ind_H^G(W) \oplus Ind_H^G(W'), \quad but \quad Ind(W \otimes W') \neq Ind_H^G(W) \otimes Ind_H^G(W').$$

On the other hand, if U is a representation of G and W is a representation of H, then:

$$\mathrm{Ind}(\mathrm{Res}(U) \otimes W) = U \otimes \mathrm{Ind}(W).$$

Indeed, $\mathrm{Ind}(W) = \bigoplus_{\sigma \in G/H} \sigma W$, so $U \otimes \mathrm{Ind}(W) = \bigoplus_{\sigma \in G/H} (U \otimes \sigma W) = \bigoplus_{G/H} \sigma(U \otimes W)$, where $U \otimes W \subset U \otimes \mathrm{Ind}(W)$ is invariant under H and isomorphic to $\mathrm{Res}(U) \otimes W$ as a H-representation. In particular:

$$\mathrm{Ind}(\mathrm{Res})(U) = U \otimes \mathrm{Ind}(\text{trivial}) = U \otimes (\text{permutation representation of } G/H).$$

We can actually calculate the character of an induced representation. Choose representatives $\sigma_1, \ldots, \sigma_k$ of cosets of H as usual. The element $g \in G$ maps $\sigma_1 W$ to $\sigma_j W$ such that $g\sigma_i \in \sigma_j H$. If $i \neq j$, this does not contribute to $\mathrm{tr}(g)$. If $i = j$, then $h = \sigma_i^{-1} g \sigma_i \in H$, and g acts on $\sigma_i W$ by $g(\sigma_i w) = \sigma_i h w$. Therefore, $\mathrm{tr}(g|_{\sigma_i W}) = \mathrm{tr}(h|_W) = \chi_W(h)$. Summing over σ_i, we get:

$$\chi_{\mathrm{Ind}(W)}(g) = \sum_{\sigma \in G/H \text{ such that } \sigma_i^{-1} g \sigma_i \in H} \chi_W(\sigma_i^{-1} g \sigma_i) = \frac{1}{|H|} \sum_{s \in G \text{ such that } s^{-1} g s \in H} \chi_W(s^{-1} g s).$$

5.10 Frobenius Reciprocity

A key property for understanding induced representations is **Frobenius reciprocity**.

Proposition 5.48. *If U is a representation of G, and W is a representation of H, then every H-equivariant map $W \to Res(U)$ extends uniquely to a G-equivariant map $Ind(W) \to U$, i.e.,*

$$Hom_H(W, Res(U)) \simeq Hom_G(Ind(W), U).$$

Proof. Choose representatives $\sigma_1, \ldots, \sigma_k \in G$ for the cosets of H, and let $V = \mathrm{Ind}(W) = \bigoplus \sigma_i W$. Given an H-equivariant map $\varphi : W \to \mathrm{Res}(U)$, if $\tilde{\varphi} : V \to U$ is G-equivariant and $\tilde{\varphi}|_W = \varphi$, we must show that $\tilde{\varphi}$ is determined uniquely.

The situation is described by the following commutative diagram:

$$\begin{array}{ccc} W & \xrightarrow{\varphi} & U \\ {\scriptstyle \sigma_i}\downarrow & & \downarrow{\scriptstyle \sigma_i} \\ \sigma_i W & \xrightarrow[\tilde{\varphi}]{} & U \end{array}$$

Thus, the map $\tilde{\varphi}|_{\sigma_i W}$ is given by

$$\tilde{\varphi}(\sigma_i w) = \sigma_i \varphi(w),$$

where $\sigma_i \in G$ acts on $\varphi(w) \in U$. This uniquely determines $\tilde{\varphi}$.

To verify that $\tilde{\varphi}$ is G-equivariant, recall that for $g \in G$, the action on V maps $\sigma_i W$ to $\sigma_j W$ where $g\sigma_i = \sigma_j h \in \sigma_j H$, and the map acts as $g(\sigma_i w) = \sigma_j hw$. Then,

$$\tilde{\varphi}(g(\sigma_i w)) = \tilde{\varphi}(\sigma_j hw) = \sigma_j \varphi(hw) = \sigma_j h\varphi(w) = g\sigma_i \varphi(w) = g(\tilde{\varphi}(\sigma_i w)).$$

This shows that $\tilde{\varphi}$ is G-equivariant on $\sigma_i W$ for all i, and hence on V.

Therefore, φ has a unique G-equivariant extension $\tilde{\varphi}$.

Conversely, given $\tilde{\varphi} \in \mathrm{Hom}_G(V, U)$, we claim that $\tilde{\varphi}$ is H-equivariant. Since $V = \bigoplus \sigma_i W$, restricting $\tilde{\varphi}$ to $W \subset V$ yields an H-equivariant map.

□

Comparing dimensions, we obtain the following corollary by noting that

$$\dim(\mathrm{Hom}_G(\dots)) = \dim(\mathrm{Hom}_H(\dots)).$$

Corollary 5.49 (Frobenius Reciprocity)**.**

$$\langle \chi_{Ind(W)}, \chi_V \rangle_G = \langle \chi_W, \chi_{Res(U)} \rangle_H.$$

Thus, if U is an irreducible representation of G and W is an irreducible representation of H, then the number of times W appears in $\mathrm{Res}(U)$ is equal to the number of times U appears in $\mathrm{Ind}(W)$.

Example 5.50. *Let $H = S_3 \subset G = S_4$. The restrictions of the irreducible representations of S_4 are:*

- *Trivial: $Res(U_4) = U_3$*
- *Alternating: $Res(U_4') = U_3'$*
- *Standard: $Res(V_4) = V_3 \oplus U_3$ (since the permutation representation $\mathbb{C}^k$ restricts to the direct sum of the permutation and trivial representations: $Res(V_4 \oplus U_4) = V_3 \oplus U_3 \oplus U_3$)*

- $V_4' = V_4 \otimes U_4'$: *Res*$(V_4') = V_3 \oplus U_3'$ *(using* $V_3 \otimes U_3' \simeq V_3$*).*
- W *(factors through* $S_4/\{(ij)(kl)\} \simeq S_3$*):* *Res*$(W) = V_3$ *(Alternatively, one can use character tables directly).*

By Frobenius reciprocity, *Ind*$(V_3) = \bigoplus$ *of the irreducible representations of* S_4 *whose restrictions contain* V_3 *(dimension 8), which gives*

$$\mathit{Ind}(V_3) = V_4 \oplus V_4' \oplus W.$$

Another example:

Example 5.51. $H = \langle (1\,2\,3\,4) \rangle \simeq \mathbb{Z}/4 \subset G = S_4$. *The irreducible representations of* H *are 1-dimensional, with* $(1\,2\,3\,4)$ *acting by powers of* i*:*

$$U_0 = \textit{trivial}, \quad U_1, U_2, U_3 : (1\,2\,3\,4) \textit{ acts by } i, i^2 = -1, i^3 = -i.$$

To find the induced representations, consider the irreducible representations of S_4 *and the eigenvalues of* $(1\,2\,3\,4)$ *and* $(1\,2\,3\,4)^2 = (1\,3)(2\,4)$*:*

$$\begin{aligned} U &\mapsto U_0, \\ U' &\mapsto U_2, \\ V &\mapsto U_1 \oplus U_2 \oplus U_3, \\ V' &\mapsto U_3 \oplus U_0 \oplus U_1, \\ W &\mapsto U_0 \oplus U_2. \end{aligned}$$

The line of reasoning is as follows: Since $\chi((1\,3)(2\,4)) = -1$*, the eigenvalues of the matrix are* $\lambda_i^2 = -1, -1, +1$*; for* $\chi(1\,2\,3\,4) = -1$*, the eigenvalues are* $\lambda_i = i, -i, -1$*. Therefore, the representation* $U \oplus V$ *has* $(1\,2\,3\,4)$ *mapped to the matrix*

$$\begin{bmatrix} 0 & & & 1 \\ 1 & \ddots & & \\ & 1 & \ddots & \\ & & 1 & 0 \end{bmatrix},$$

with eigenvalues $\pm 1, \pm i$*, and we find that* *Res*$(U \oplus V) = U_0 \oplus \cdots \oplus U_3$*.*

Frobenius reciprocity implies the following induced representations:

$$\begin{aligned} &\mathit{Ind}(U_0) = U \oplus V' \oplus W \quad (\textit{permutation representation of } S_4 \textit{ on } G/H), \\ &\mathit{Ind}(U_1) = \mathit{Ind}(U_3) = V \oplus V', \\ &\mathit{Ind}(U_2) = U' \oplus V \oplus W \quad (\simeq U' \otimes \mathit{Ind}(U_0), \textit{ consistent with } U_2 = \mathit{Res}(U')). \end{aligned}$$

Some of the key motivation for studying induced representations comes from two deep theorems of Artin and Brauer:

Theorem 5.52 (Artin)**.** *Every character of a representation of G is a linear combination with rational coefficients of characters of representations induced from cyclic subgroups of G.*

Theorem 5.53 (Brauer)**.** *Every character of a representation of G is a linear combination with integer coefficients of characters of representations induced from "elementary" subgroups of G, where "elementary" means isomorphic to a product $C \times H$, with H a p-group ($|H| = p^k$) and C cyclic ($C \simeq \mathbb{Z}/n$, $p \nmid n$).*

We won't cover the proofs of these theorems in this course.

5.11 Group Algebra

The group algebra of a finite group G provides an alternative perspective on the representations of G. Although it is not as immediately useful for calculating characters and finding irreducibles, it is conceptually important.

Definition 5.54. *The **group algebra** of G is the vector space $\mathbb{C}G = \left\{\sum_{g\in G} a_g e_g \mid a_g \in \mathbb{C}\right\}$, with the product $e_g \circ e_h = e_{gh}$ (extended by linearity). This is a noncommutative ring, with multiplication given by:*

$$\left(\sum_g a_g e_g\right)\left(\sum_g b_g e_g\right) = \sum_g \left(\sum_h a_h b_{h^{-1}g}\right) e_g$$

(commutative if and only if G is abelian).

As a vector space, $\mathbb{C}G$ is isomorphic to the regular representation of G; the novelty here is in the multiplication structure.

An action of G on a vector space V (a representation) is a homomorphism $\rho : G \to \mathrm{GL}(V)$, which extends linearly to an algebra homomorphism (i.e., a linear map of vector spaces that preserves multiplication) $\mathbb{C}G \to \mathrm{End}(V)$. This map sends basis elements $e_g \mapsto \rho(g)$ and extends linearly: $\sum a_g e_g \mapsto \sum a_g \rho(g)$. To verify that this map is compatible with multiplication, we use (bi)linearity. It is enough to check for basis elements: $e_g \cdot e_h = e_{gh} \mapsto \rho(gh) = \rho(g) \circ \rho(h)$.

Proposition 5.55. *A G-representation is the same as a (left) $\mathbb{C}G$-module. That is, a vector space V and an action $\mathbb{C}G \times V \to V$, given by a ring homomorphism $\mathbb{C}G \to End(V)$.*

Example 5.56. *The regular representation of G corresponds to $\mathbb{C}G$ as a module over itself, where the operation of $\mathbb{C}G$ is left-multiplication.*

Since we haven't delved deeply into rings and modules, we will not explore this further. However, there is one elegant result worth mentioning:

Given a finite group G, let $V_1, \ldots, V_r$ be the irreducible representations of G. Each of these representations gives a ring homomorphism $\mathbb{C}G \to \mathrm{End}(V_i)$. Together, these yield a map $\mathbb{C}G \to \bigoplus_{k=1}^r \mathrm{End}(V_i)$ (which is a subring of $\mathrm{End}\left(\bigoplus_{i=1}^r V_i\right)$—the subring of block-diagonal linear operators on $\bigoplus_{i=1}^r V_i$). This map is again a

ring homomorphism, with the product in $\mathbb{C}G$ mapping to the composition in $\mathrm{End}(V_i)$.

Proposition 5.57. *If $V_1, \ldots, V_r$ are the irreducible representations of G, then the map $\mathbb{C}G \to \bigoplus_{i=1}^r End(V_i)$ is an isomorphism of rings.*

Proof. We already know that the map is a homomorphism, so we just need to verify that it is bijective.

- *Injectivity*: Assume that $\sum a_g e_g \in \mathbb{C}G$ belongs to the kernel. Then for all irreducible representations, we have $\sum a_g \rho_{V_i}(g) = 0$, implying that for all representations of G, $\sum a_g \rho(g) = 0$. However, for the regular representation, the maps $\rho(g)$ are linearly independent (since $\sum a_g \rho_g$ maps e_1 to $\sum a_g e_g$). This implies that $a_g = 0$ for all g.
- *Surjectivity*: We have $\dim(\mathbb{C}G) = |G| = \sum(\dim(V_i))^2 = \dim\left(\bigoplus_{i=1}^r \mathrm{End}(V_i)\right)$, so an injective linear map is surjective.

□

In the ring $\bigoplus \mathrm{End}(V_i)$, as in any direct sum of rings, the projectors onto each summand are given by:

$$P_i = \begin{cases} \mathrm{Id} & \text{on } \mathrm{End}(V_i), \\ 0 & \text{on } \mathrm{End}(V_j), j \neq i \end{cases}$$

These projectors are orthogonal idempotents: $P_i^2 = P_i$ and $P_i P_j = 0$ for $i \neq j$.

By comparison with projection formulas: we have seen that for all representations V, the map $\varphi_i = \frac{\dim(V_i)}{|G|} \sum_g \overline{\chi_{V_i}(g)} g : V \to V$ is the projection onto the V_i-summands. This means the idempotents of $\mathbb{C}G$ corresponding to the projectors P_i under the isomorphism are:

$$\pi_i = \frac{\dim(V_i)}{|G|} \sum_{g \in G} \overline{\chi_{V_i}(g)} e_g \in \mathbb{C}G$$

The identities $\pi_i^2 = \pi_i$ and $\pi_i \pi_j = 0$ for $i \neq j$ recover, among other things, the orthonormality of the characters χ_{V_i}. Given a $\mathbb{C}G$-module V, it has a submodule $\pi_i V$—these are the pieces of V corresponding to the V_i-summands in the decomposition of V.

5.12 Real Representations

We have studied actions of finite groups on complex vector spaces, and now we want to extend this study to real vector spaces.

If V_0 is a representation of G over $\mathbb{R}$, it has an **invariant inner product** $\langle \cdot, \cdot \rangle$. This can be constructed starting from any inner product $b(\cdot, \cdot)$ and defining

$$\langle v_1, v_2 \rangle = \frac{1}{|G|} \sum_{g \in G} b(gv_1, gv_2).$$

This ensures that the elements of G act by orthogonal transformations (isometries).

This leads to **complete reducibility**: every representation over $\mathbb{R}$ splits into the direct sum of irreducibles (the same proof as in the complex case: if $U_0 \subset V_0$ is an invariant subspace (subrepresentation), then $V_0 = U_0 \oplus U_0^{\perp}$).

However, Schur's Lemma does not hold in the same way:

Example 5.58. *Consider the action of $\mathbb{Z}/n$ on $\mathbb{R}^2$ by rotations, where k acts by the matrix*

$$\begin{bmatrix} \cos\left(\frac{2\pi k}{n}\right) & -\sin\left(\frac{2\pi k}{n}\right) \\ \sin\left(\frac{2\pi k}{n}\right) & \cos\left(\frac{2\pi k}{n}\right) \end{bmatrix}.$$

This is irreducible as a representation over $\mathbb{R}$, but the representation has automorphisms that are not multiples of Id: any rotation of $\mathbb{R}^2$ is $\mathbb{Z}/n$-equivariant.

Therefore, many of the results we developed for complex representations do not directly apply to real representations. Instead, we must use the concept of *complexification*, which we have already encountered when studying operators on real vector spaces.

We define a map

$$\{\text{real representations of } G\} \to \{\text{complex representations of } G\}, \quad V_0 \mapsto V_0 \otimes_{\mathbb{R}} \mathbb{C} = V_0 \oplus iV_0,$$

where G acts on $V = V_0 \oplus iV_0$ by

$$g(v + iw) = gv + igw.$$

In other words, given a basis (e_j) of V_0, we have $e_j + 0i$ (which is just e_j) as a basis of V, and G acts by the same matrix on both V_0 and V.

Definition 5.59. *A complex representation V of G is called **real** if there exists a representation V_0 over $\mathbb{R}$ such that $V = V_0 \otimes_{\mathbb{R}} \mathbb{C}$.*

A necessary condition for V to be real is that its character χ_V must take real values, because the matrix $g : V \to V$ in a suitable basis has real entries. However, this is not a sufficient condition.

Example 5.60. *Consider the quaternion group $Q = \{\pm 1, \pm i, \pm j, \pm k\}$, with the relations $i^2 = j^2 = k^2 = ijk = -1$. The group Q acts by*

$$\pm 1 \mapsto \pm Id, \quad \pm i \mapsto \pm \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}, \quad \pm j \mapsto \pm \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad \pm k \mapsto \pm \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}.$$

The character $\chi(\pm 1) = \pm 2$ and $\chi(\pm i) = \chi(\pm j) = \chi(\pm k) = 0$, so χ takes real values.

However, this does not come from a 2-dimensional real representation of Q in $GL(2, \mathbb{R})$. If there were such a representation, it would have an invariant inner product, so Q would embed in $O(2)$, with -1 acting by $-Id$. But only two elements of $O(2)$ square to $-Id$ (the rotations by $\pm 90°$), whereas we need four such elements for $\pm i, \pm j, \pm k$.

To classify representations over $\mathbb{R}$ using characters, we need to understand which complex representations are real. We will focus on this for irreducible representations over $\mathbb{C}$. However, note that if V_0 is an irreducible representation over $\mathbb{R}$, then $V = V_0 \otimes_{\mathbb{R}} \mathbb{C}$ can still be reducible over $\mathbb{C}$ (for example, the rotations of $\mathbb{R}^2$ by $\mathbb{Z}/n$).

Proposition 5.61. *A complex representation V of G is real if and only if there exists a G-equivariant complex antilinear map $\tau : V \to V$ (i.e., $\tau(\lambda v) = \overline{\lambda}\tau(v)$) such that $\tau^2 = id$.*

Proof. One direction is clear: if $V = V_0 \otimes_{\mathbb{R}} \mathbb{C}$, define $\tau(v + iw) = v - iw$ for $v, w \in V_0$, which corresponds to complex conjugation.

For the opposite direction, given such a map τ, decompose any $v \in V$ as $\mathrm{Re}(v) = \frac{v+\tau(v)}{2}$ and $i\mathrm{Im}(v) = \frac{v-\tau(v)}{2}$, which belongs to the ± 1-eigenspaces of τ. Let $V_0 = \mathrm{Ker}(\tau - \mathrm{id})$, which is an $\mathbb{R}$-subspace of V (but not a $\mathbb{C}$-subspace). Since τ is $\mathbb{R}$-linear, we have $\tau i = -i\tau$, so iV_0 is the -1-eigenspace, and $V = V_0 \oplus iV_0 \simeq V_0 \otimes_{\mathbb{R}} \mathbb{C}$.

Since τ is G-equivariant, the eigenspaces $V_0 = \mathrm{Ker}(\tau - \mathrm{id})$ and iV_0 are preserved by G, and thus both are subrepresentations over $\mathbb{R}$.

□

Now, let V be an irreducible complex representation of G such that χ_V takes real values. In this case, $\chi_V = \overline{\chi_V} = \chi_{V^*}$, so $V \simeq V^*$ as G-representations.

Let $\varphi : V \xrightarrow{\sim} V^*$ be such an isomorphism (which, by Schur's Lemma, is unique up to multiplication by a scalar $\lambda \in \mathbb{C}^*$).

Recall that a linear map $\varphi : V \to V^*$ determines a bilinear form $B : V \times V \to \mathbb{C}$, where $B(v, w) = \varphi(v)(w)$. The invariance of B under the action of G implies that φ is G-equivariant. Thus, V admits a G-invariant bilinear form B, unique up to scaling, and nondegenerate if nonzero.

Recall that $B \in (V \otimes V)^* = \mathrm{Sym}^2(V^*) \oplus \wedge^2 V^*$, i.e., the symmetric and skew parts of B. By uniqueness, one of these parts is zero and the other is nondegenerate. Thus, B is either symmetric or skew-symmetric.

The symmetric case corresponds to real representations, while the skew-symmetric case corresponds to quaternionic representations.

5.13 Quaternionic Representations

Proposition 5.62. *An irreducible complex representation V of a finite group G is real if and only if V contains a G-invariant nondegenerate symmetric bilinear form $B : V \times V \to \mathbb{C}$.*

Proof. Assume $V = V_0 \otimes_{\mathbb{R}} \mathbb{C}$ is real. Then V_0 has an invariant real inner product B; extend it to a $\mathbb{C}$-bilinear form as follows:

$$B(v_1 + iw_1, v_2 + iw_2) := B(v_1, v_2) + iB(w_1, v_2) + iB(v_2, w_1) - B(v_2, w_2),$$

which defines a nondegenerate symmetric bilinear form on V.

Conversely, if $B : V \times V \to \mathbb{C}$ determines an isomorphism $\varphi : V \to V^*$ (which is $\mathbb{C}$-linear and equivariant), choosing an invariant Hermitian inner product H on V, we also obtain a $\mathbb{C}$-antilinear equivariant bijection $V \to V^*$. Composing the two maps gives a $\mathbb{C}$-antilinear equivariant map $\tau : V \to V$, characterized by

$$H(\tau(v), w) = B(v, w).$$

Now, τ^2 is an equivariant $\mathbb{C}$-linear isomorphism $V \to V$, so by Schur's lemma, $\tau^2 = \mathrm{Id}$. A calculation shows:

$$H(\tau^2(v), v) = B(\tau(v), v) = B(v, \tau(v)) = H(\tau(v), \tau(v)) \geq 0,$$

which implies that $\lambda \in \mathbb{R}_+$. By rescaling H as $\lambda^{1/2}H$, we can arrange that $\tau^2 = \mathrm{id}$. Thus, V is real by the previous proposition. □

In the case where the invariant bilinear form B is skew-symmetric, the same argument yields a $\mathbb{C}$-antilinear equivariant bijective map $J : V \to V$, which now satisfies $J^2 = -\mathrm{id}$. This defines a **quaternionic** structure on V, i.e., it describes an $\mathbb{H}$-module structure on V, where $\mathbb{H}$ is the quaternions:

$$\mathbb{H} = \{a + bi + cj + dk \mid a, b, c, d \in \mathbb{R}\}, \quad i^2 = j^2 = k^2 = ijk = -1,$$

a division algebra (noncommutative analogue of a field: $\mathbb{H}$ is a noncommutative ring such that every nonzero element has a multiplicative inverse). We have $\mathbb{H} = \mathbb{C}1 \oplus \mathbb{C}j$, with the relations $ji = -ij$ and $j^2 = -1$. Thus, an $\mathbb{H}$-module is equivalent to a $\mathbb{C}$-vector space, together with an antilinear map j such that $j^2 = -\mathrm{id}$.

Example 5.63. *The regular representation V of S_3 is real. This can be seen directly by noting that $S_3 \simeq D_3$ acts on $V_0 = \mathbb{R}^2$ by rotations and reflections, and $V_0 \otimes_{\mathbb{R}} \mathbb{C} \simeq V$. Alternatively, one can observe that $V^* \simeq V$, and $\wedge^2(V^*) \simeq U'$ has no trivial summand, hence there is no invariant skew-symmetric bilinear form $B \in \wedge^2 V^*$. However, $\mathrm{Sym}^2(V^*)$ does have such a form, which can be applied using the argument above.*

Example 5.64. *The quaternion group is a quaternionic representation:* $\mathbb{H} \simeq \mathbb{C} \oplus j\mathbb{C}$. *The linear maps correspond to left multiplication by elements of* Q *(e.g.,* $i(z_1 + jz_2) = iz_1 + j(-iz_2)$ *and* $k(z_1 + jz_2) = -iz_2 + j(-iz_1)$*). The* $\mathbb{C}$*-antilinear map* $J : V \to V$*, where* $J^2 = -1$*, is right multiplication by* j*, and it commutes with left multiplication by* v*.*

6 Point Set Topology

6.1 Metric Spaces

What is topolog? Unlike geometry, which concerns quantitative information about spaces (distances, volumes,...), topology concerns itself with qualitative properties that are invariant under continuous deformation. Eg: is it connected (a single piece) simply connected? (sphere vs torus)

Point-set topology also gives a language (topological spaces, open and closed sets, compactness) but for algebraic topology (associate algebraic invariants to spaces, eg fundamental group) and for analysis.

Example 6.1 (Extreme Value Theorem)**.** *$f : [a, b] \to \mathbb{R}$ continuous $\implies$ f achieves it max and min at some points of $[a, b]$.*

This is in fact true for any continuous $f : X \to \mathbb{R}$ whenever X is a compact topological space, and is a special instance of:

Theorem 6.2. *If $f : X \to Y$ continuous mapping between topological spaces and X is compact, then $f(X)$ is compact.*

Since the general notion of topological space is quite abstract; let's start with a more familiar class of examples: metric spaces.

Example 6.3. *A metric space (X, d) is a set X together with a **distance function** $d : X \times X \to \mathbb{R}_{\geq 0}$ such that*

1. *For $p, q \in X$, $d(p, q) = 0 \leftrightarrow p = q$*
2. *For $p, q \in X$, $d(p, q) = d(q, p)$*
3. *For $p, q, r \in X$, $d(p, r) \leq d(p, q) + d(q, r)$ (the triangle inequality)*

Example 6.4. *$X = \mathbb{R}^n$ with Euclidean distance $d(x, y) = \left(\sum_{i=1}^n (y_i - x_i)^2\right)^{\frac{1}{2}}$*

Example 6.5. *If $Y \subset X$ then $(Y, d|_Y)$ is a metric space (induced metric)*

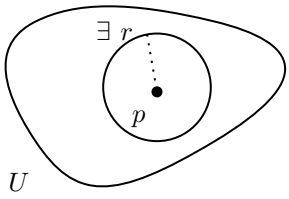
Example 6.6. *Different metrics on $\mathbb{R}^n$: $d_1(x, y) = \sum_{i=1}^n |y_i - x_i|$, $d_\infty(x, y) = max(|y_i - x_i|)$*

Exercise: check $(\mathbb{R}^n, d_1)$ and $(\mathbb{R}^n, d_\infty)$ are metric spaces. What do balls look like?

Definition 6.7. *Let (X, d) be a metric space, $p \in X, r > 0$: the **open ball** of radius r around p is $B_r(p) = \{q \in X | d(p, q) < r\}$.*

Here is a more general notion:

Definition 6.8. *$U \subset X$ is open if $\forall p \in U, \exists r > 0$ s.t. $B_r(p) \subset U$.*

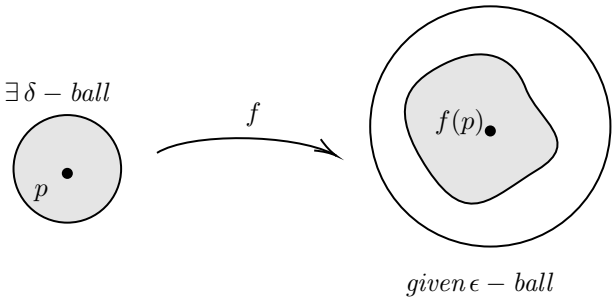


Proposition 6.9. *Open balls are open; so are arbitrary unions and finite intersections of open sets.*

In fact, open sets are unions of open balls! $U = \bigcup_{p \in U} B_{r(p)}(p)$.

This is useful to a general discussion of continuity:

Definition 6.10. $(X, d_X), (Y, d_Y)$ *metric spaces.* $f : X \to Y$ *is* ***continuous*** *if* $\forall p \in X, \forall \epsilon > 0, \exists \delta > 0$ *s.t.* $d_X(p, x) < \delta \implies d_Y(f(p), f(x)) < \epsilon$.



Theorem 6.11. $f : X \to Y$ *is continuous if and only if* $\forall U \subset Y$ *open,* $f^{-1}(U) \subset X$ *is open.*

Proof. Assume f is continuous. Let $U \subset Y$ open, let $p \in f^{-1}(U)$, i.e. $f(p) \in U$. Since U is open, $\exists \epsilon > 0$ such that $B_\epsilon(f(p)) \subset U$. By continuity, $\exists \delta > 0$ such that $d_X(p, x) < \delta \implies f(x) \in B_\epsilon(f(p)) \subset U$. Hence $B_\delta(p) \subset f^{-1}(U)$. So $f^{-1}(U)$ is open.

Conversely, assume U open $\implies f^{-1}(U)$ is open. Fix $p \in X$, $\epsilon > 0$. $B_\epsilon(f(p))$ is open in Y, so $f^{-1}(B_\epsilon(f(p))) \ni p$ is open in X. Hence $\exists \delta > 0$ such that $B_\delta(p) \subset f^{-1}(B_\epsilon(f(p)))$. This means $d_X(p, x) < \delta \implies x \in f^{-1}(B_\epsilon(f(p))) \implies f(x) \in B_\epsilon(f(p))$. □

We cann also talk about sequences and their limits:

Definition 6.12. *A sequence* $p_1, p_2, \ldots$ *in* (X, d) *converges to a* ***limit*** $p \in X$ *(write* $p_n \to p$ *or* $\lim_{n\to\infty} p_n = p$*) if* $\forall \epsilon > 0, \exists N$ *s.t.* $\forall n \geq N, d(p_n, p) < \epsilon$.

Remark 6.13. *The limit is unique if it exists.*

Definition 6.14. *A sequence* $p_1, p_2, \ldots$ *in* X *is* ***Cauchy*** *if* $\forall \epsilon > 0$, $\exists N$ *s.t.* $\forall m, n \geq N, d(p_n, p_m) < \epsilon$.

The difference bs that p_n get closer to p v.s. get closer to each other.

Exercise 6.15. *If a sequence converges then it is Cauchy, but not necessarily vice-versa.*

Definition 6.16. *A metric space is* ***complete*** *if every Cauchy sequence converges.*

Example 6.17. $\mathbb{R}$ *is complete, but* $\mathbb{Q}$ *(with the induced metric) isn't complete.*

The notion of Cauchy sequence is sepcific to metric spaces, but reallyt useful for real analsysis.

Example 6.18. $e = \sum_{k=0}^{\infty} \frac{1}{k!}$ *- if we take this to be th edefinition of e, we can't prove direct that* $x_n = \sum_{k=0}^{n} \frac{1}{j!}$ *converges to e, instead use Cauchy criterion to show that the liit exists.*

Interlude:what is $\mathbb{R}$?

It's an ordered field (ie: $+-\times/$ and $<$ compatible with usual rules) with the least upper bound property: every nonempty subset $E \subset \mathbb{R}$ that admits an upper bound ($\exists M \in \mathbb{R}$ s.t. $\forall x \in E, x \leq M$) has a least upper bound $\sup(E) \in \mathbb{R}$ (ie $\sup(E)$ is an upper bound, and every upper bound for E is $\geq \sup(E)$).

The least upper bound property is equivalent to completeness of $\mathbb{R}$; any ordered field with this property is isomorphic to $(\mathbb{R}, +, \times, <)$. Constructions of $\mathbb{R}$ from $\mathbb{Q}$ involve adding thee missing elements (irrationals) so that the least upper bound property and completeness holds; the elements of $\mathbb{R}$ end up being either the sups of certain subsets of $\mathbb{Q}$ or the limits of Cauchy sequences of $\mathbb{Q}$.

Returning to limits of sequences...

Proposition 6.19. *If* $p_n \to p$*, then every open subset* $U \ni p$ *contains* p_n *for all but finitely many* n*.*

This will be the definition of limit outside the metric case.

Proof. $U \ni p$, U open $\implies$ $\exists \epsilon > 0$ s.t. $B_\epsilon(p) \in U$. So $\exists N$ s.t. $n \geq N \implies p_n \in B_\epsilon(p) \subset U$. □

Definition 6.20. $Z \subset X$ *is* ***closed*** *if its complement* $X \backslash Z$ *is open.*

Most subsets of X are neither open nor closed... and $\emptyset$ and X are both!

Proposition 6.21. *If* $Z \subset X$ *is closed, then* $\forall$ *sequence* $\{p_n\}$ *in* Z *which converges to a limit* $p \in X$*, then* $p \in Z$*.*

The converse is true in metric spaces and in nice enough topological spaces - first countable.

Proof. Assume $\exists \{p_n\} \in Z$, $p \in X \backslash Z$, $p_n \to p$: $\forall U \ni p$ open, U contains p_n for all but finitely many n, but $p_n \in Z$, so $U \not\subset X \backslash Z$. If Z is closed then $U = X \backslash Z$ is open and we get a contradicgion. □

Our goal will be to reformulate/generalize al this in the context of topoligcal spaces, i.e. sets equipped with a topology which may or may not come from a matric.

6.2 Topological Spaces

We will now reformulate/generalize all this in the context of topological spaces, ie. sets equipped with a topology which may or may not come from a metric.

Definition 6.22. *A **topological space** is a set X together with a collection $\tau \subset P(X)$, the **open sets** in X, such that*

- $\emptyset \in \tau, X \in \tau$
- *arbitrary unions of open sets are open.*
- *finite intersections of open sets are open.*

Why bother? One answer: many natural topologies do not come from a metric! Eg, in analysis:

- On the space of (bounded) functions $f : X \to \mathbb{R}$, uniform convergence topology ($f_n \to f$ if and only if $\sup_x |f_n(x) - f(x)| \to 0$) comes from a metric ($d(f,g) = \sup_x |f(x) - g(x)|$) but pointwise convergence ($f_n \to f$ if and only if $\forall x \in X$, $f_n(x) \to f(x)$) doesn't."product topology")
- C^∞ topology on smooth functions $\mathbb{R} \to \mathbb{R}$ doesn't come from a metric either.

And on the other hand, a metric contains extraneous information for topology. $(\mathbb{R}^n, d), (\mathbb{R}^n, d_1), (\mathbb{R}^n, d_\infty)$ have the same open sets $\implies$ same topologies.

Definition 6.23. *A function $f : X \to Y$ is **continuous** if $\forall U \subset Y, U$ open $\implies f^{-1}(U) \subset X$ is open.*

Definition 6.24. *A sequence $\{p_n\}$ in X converges to a limit p ($p_n \to p$) if $\forall U \ni p$ open, $\exists N \in \mathbb{N}$ such that $n \geq N \implies p_n \in U$.*

Example 6.25.

- *(X,d) metric space $\implies$ $\tau = \{U \subset X | \forall p \in U \exists \epsilon > 0$ such that $B_\epsilon(p) \subset U\}$ metric topology.*
- *Discrete topology: $\tau = \mathcal{P}(X)$ (every subset is open and closed.) eg. usual topology on $\mathbb{Z} \subset \mathbb{R}$, which is in fact a metric topology: set $d(x,y) = 1 \forall x \neq y$.*

These abstract definitions imply basic facts about continuity, such as:

Proposition 6.26.

- *If $f : X \to Y$ continuous, $p_n \to p$ in $X \implies f(p_n) \to f(p)$ in Y.*
- *$f : X \to Y$ and $g : Y \to Z$ continuous $\implies g \circ f : X \to Z$ continuous.*

Given two topologies τ, τ' on X, if $\tau \subset \tau'$ we say τ' is finer and τ is coarser. The finest topology on X is the discrete one (all points are isolated), while the coarsest is $\{\emptyset, X\}$ ("one big clump").

The finer topology τ' has more open sets; it's easier for functions $X \to Y$ to be continuous with respect to τ than τ (every function from a discrete set is continuous). It's harder for sequences to converge in τ' (eg. on a discrete set, convergent subsequences must be constant after finitely many terms; while for $\tau = \{\emptyset, X\}$, every sequence converges to every point of X, in particular limit isn't unique!)

6.3 Bases

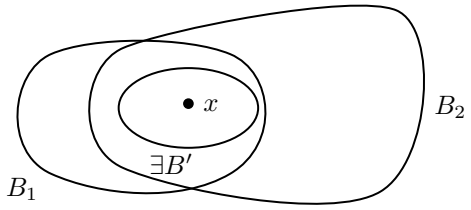
Keeping track of all the open sets is cumbersome - in metric spaces we started with open balls and got a characterization of open sets in terms of these.

The analogous notion for a general topology is that of basis:

Definition 6.27. *Assume $\mathcal{B} \subset \mathcal{P}(X)$ is a collection of subsets of X such that*

1. $\bigcup_{B \in \mathcal{B}} B = X$
2. *If $B_1, B_2 \in \mathcal{B}$ and $x \in B_1 \cap B_2$ then $\exists B' \in \mathcal{B}$ such that $x \in B' \subset B_1 \cap B_2$.*

*Then we say $\mathcal{B}$ is a **basis** and **generates** the topology $\tau =$ arbitrary unions of elements of $\mathcal{B}$. Equivalently: $U \in \tau \iff \forall x \in U \exists B \in \mathcal{B}$ such that $x \in B \subset U$.*



Check: the two characterizations of τ are equivalent, and τ is a topology.

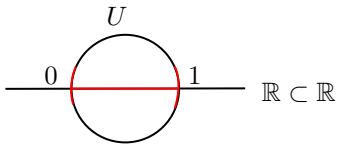
Remark 6.28. *Unlike bases in linear algebra, bases in topology can contain redundant info - a better analogy is with generating sets... eg. metric topology is generated by any of: all open sets; open balls $\mathcal{B}_r(x), x \in X, r > 0$; open balls $\mathcal{B}_{1/n}(x), x \in X$; open balls $\mathcal{B}_{1/n}(y), y \in Y \subset X$ dense subset (every nonempty open intersects Y) eg. $\mathbb{Q} \subset \mathbb{R}$. So for example the usual topology on $\mathbb{R}$ on $\mathbb{R}^n$ actually admits a countable basis.*

6.4 Subspaces and Products

How do you make new topological spaces? Subspaces and products.

Definition 6.29. *(X, τ_X) topological space, $Y \subset X$ any subset $\implies$ the **subspace topology** on Y is $\tau_U = \{U \cap Y | U \in \tau_X\}$.*

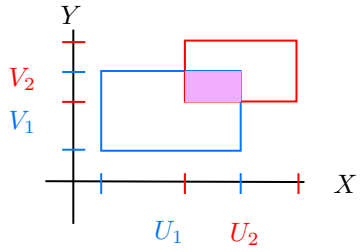
It's important when stating "U is open" to be clear: as a subset of what space? Eg. Y is always open as a subset of itself! $(0,1) \subset \mathbb{R} \subset \mathbb{R}^2$ is open in $\mathbb{R}$ but not in $\mathbb{R}^2$.



It's the coarsest topology on Y that makes the inclusion $Y \hookrightarrow X$ continuous. Also, if τ_X comes from a metric d on X, then τ_Y comes from $d|_Y : Y \times Y \hookrightarrow X \times X \xrightarrow{d} R_{\geq 0}$.

Definition 6.30. *$(X, \tau_X), (Y, \tau_Y)$ topological spaces $\implies$ the **product topology** on $X \times Y$ is the topology generated by basis $\mathcal{B} = \{U \times V | U \subset X$ open, $V \subset Y$ open$\}$.*

When X, Y are metric spaces, this is also a metric topology, defined by $d^{\infty}_{X\times Y}((x_1,y_1)(x_2,y_2)) = \max(d_X(x_1,x_2) + d_Y(y_1,y_2))$. Or in fact $d^2_{X\times Y} = \sqrt{d_X(x_1,x_2)^2 + d_Y(y_1,y_2)^2}$, $d^1_{X\times Y} = d_X(x_1,x_2) + d_Y(y_1,y_2)$ define the same topology on $X \times Y$. So this gives the usual topology on $\mathbb{R}^n$.



In general, it's the coarsest topology on $X \times Y$ such that the projection maps $X \times Y \xrightarrow{p_1} X, X \times Y \xrightarrow{p_2} Y$ are continuous. Also: $(x_n, y_n) \to (x, y)$ if and only if $x_n \to x$ and $y_n \to y$.

Similarly for finite products $X_1 \times ... \times X_n$. For infinite products these are several different natural topologies; we'll see this later.

Homeomorphisms: what is the correct notion of 2 topological spaces being "the same"?

Definition 6.31. *X, Y are **homeomorphic** if there exist continuous maps $f : X \to Y$ and $g : Y \to X$ such that $f \circ g = id_Y, g \circ f = id_X$.*

Equivalent, a homeomorphism $f : X \to Y$ is a continuous bijection such that f^{-1} continuous, ie: a bijection $X \iff Y$ under which $\tau_X \iff \tau_Y$.

Remark 6.32.

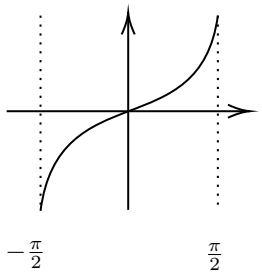
- *A continuous bijection need not be a homeomorphism. Eg. X with 2 topologies, τ' strictly finer than $\tau \implies (X, \tau') \to (X, \tau)$ is a bijection,*

continuous since $U \in \tau \implies id^{-1}(U) = U \in \tau'$ *but not homeomorphic.*

- *Say a metric space* (X, d) *is* ***bounded*** *if* $diam(X) = sup\{d(p,q)|p,q \in X\} < \infty$. *This is not a topological property, eg.*

$$f : (-\frac{\pi}{2}, \frac{\pi}{2}) \to \mathbb{R},$$
$$x \mapsto \tan x$$

is a homeomorphism (tan and arctan are continuous), so $\mathbb{R}$ *is homeomorphic to* $(-\frac{\pi}{2}, \frac{\pi}{2})$ *(or any open interval in* $\mathbb{R}$*).*



6.5 Interior and Closure

Definition 6.33. *A subset* A *of a topological space* X *is* ***closed*** *if* $X \backslash A$ *is open.*

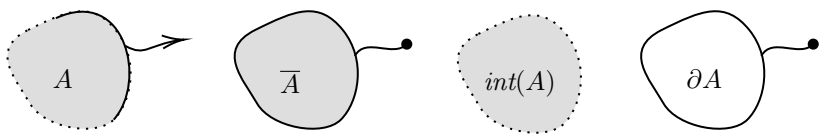
Remark 6.34. *Subsets can be both closed and open, eg.* $\emptyset$ *and* X*, or neither (eg.* $[0, 1)$ *or* $\mathbb{Q}$ *in* $\mathbb{R}$*).*

Axioms of open sets imply:

- $\emptyset, X$ are closed.
- arbitrary intersection of closed sets are closed.
- finite unions of closed sets are closed.

Definition 6.35. *For* $A \subset X$ *any subset, we define:*

1. *The* ***closure*** *of* A*:* $\bar{A}$ = *smallest closed set containing* $A = \bigcap_{A \subset F, F \text{ closed}} F$ *(*$A \subset \bar{A}$*,* $\bar{A}$ *closed since it's* $\cap$ *of closed)*
2. *The* ***interior*** *of* A*,* $int(A)$ = *largest open set contained in* $A = \bigcup_{U \subset A, U \text{ open}} U$ *(open).*
3. *The* ***boundary*** *of* A *is* $\partial A = \bar{A} - int(A)$.

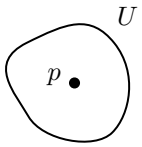


Example 6.36. *$A = [0,1) \subset \mathbb{R}$, usual topology $\implies$ $\bar{A} = [0,1], int(A) = (0,1), \partial A = \{0,1\}$.*

Remark 6.37.

- *A is closed if and only if $\bar{A} = A$, open if and only if $int(A) = A$.*
- *$\overline{X \backslash A} = X - int(A), int(X - A) = X - \overline{A}$.*

Definition 6.38. *Say $U \subset X$ is a **neighborhood** of $p \in X$ if U is open and $p \in U$.*



Proposition 6.39.

1. *$p \in int(A)$ if and only if A contains a neighborhood of p.*
2. *$p \in \bar{A}$ if and only if every neighborhood of p intersects A nontrivially.*

Proof.

1. $p \in \text{int}(A) \iff \exists U$ open such that $p \in U \subset A$.
2. $p \in A \iff p \notin \text{int}(X - A) \iff \forall U \ni \text{open}, A \cap U = \emptyset$.

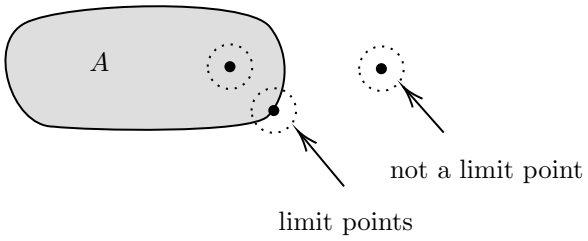
□

Definition 6.40. *Say A is **dense** if $\bar{A} = X$ (ie. every nonempty open subset of X intersects A nontrivially).*

Example 6.41. *$\mathbb{Q}$ is dense in $\mathbb{R}$ (for usual topology).*

6.6 Closed Sets and Limit Points

Definition 6.42. *$x \in X$ is a **limit point** of $A \subset X$ if, for every neighborhood $U \ni x$, $U \cap (A - \{x\}) \neq \emptyset$.*



Example 6.43. *In $\mathbb{R}_{std}$, 1 is a limit point of $(0,1)$ and of $[0,1]$, but is not a limit point of $\{\frac{1}{n}, n \geq 1\} \cap \{0\}$ (but 0 is).*

Proposition 6.44. $\bar{A} = A \cup \{\textit{limit points of } A\}$.

Proof. $A \subset \bar{A}$ by definition, so it's enough to consider points not in A. If $x \notin A, \forall U \ni x$ neighborhood, $U \cap (A - \{x\}) = U \cap A$, so $x \in \bar{A}$ if and only if x limit point.

□

Corollary 6.45. *A is closed if and only if A contains all of its limit points.*

What is the connection between limit points and limits of sequences?

Proposition 6.46. *$p \in X$:*

- *if $\exists \{p_n\}$ sequence in $A \subset X$ such that $p_n \to p$ then $p \in \bar{A}$.*
- *if $\exists \{p_n\}$ sequence in $A, p_n \neq p$ for ∞ many $n, p_n \to p$, then p is a limit point of A.*

Proof. Any neighborhood $U \ni p$ contains p_n for all large n, hence contains points of A.

□

The converse is true in metric spaces: if $p \in \bar{A}$ (resp. a limit point of A) then $\forall n > 0, \exists p_n \in B_{1/n}(p) \cap A$ (resp. with $p_n \neq p$), so $\exists$ sequence in A such that $p_n \to p$.

This holds more generally in spaces whose points have countable bases of neighborhoods $U_1 \subset U_2 \subset ...$ (ie. $\forall p, \exists$ neighborhoods $U_1, U_2, ...$ such that $\forall$ neighborhoods $U \ni p, \exists n$ such that $p \in U_n \subset U$), but not in arbitrary topological spaces.

6.7 Hausdorff Spaces

In a metric space, a sequence converges to at most one limit. This is not true in an arbitrary topological space!

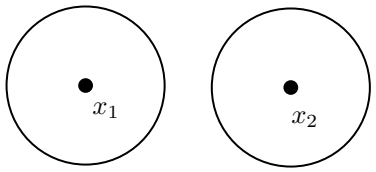
Example 6.47. *$X = \mathbb{R}$ with finite complement topology: open subsets $= \emptyset$ and $\mathbb{R} - \{\textit{finite subsets}\}$. Let $a_1, ...$ be a sequence in X with all a_i distinct. Then $\forall x \in X$, every neighborhood $U \ni x$ contains all but finitely many of the a_i, hence $\exists N$ such that $a_n \in U \forall n \geq N$. Thus the sequence converges to every point of X.*

To avoid such pathological behavior:

Definition 6.48. *A topological space is **Hausdorff** (or T_2) if $\forall x_1 \neq x_2 \in X$, $\exists$ neighborhoods $U_1 \ni x_1, U_2 \ni x_2$ such that $U_1 \cap U_2 = \emptyset$.*

Example 6.49.

1. *Any metric space is Hausdorff: given $x_1 \neq x_2$, choose $0 \leq \epsilon < \frac{1}{2}d(x_1, x_2)$ then $U - i = B_\epsilon(x_i)$ disjoint neighborhoods of x_i.*



$B_\epsilon(x_1)$ $B_\epsilon(x_2)$

2. *The finite complement topology on $\mathbb{R}$ is not Hausdorff, since any two non=empty open sets intersect (in infinitely many points).*
3. *The discrete topology is always Hausdorff ($U_i = \{x_i\}$ disjoint neighborhoods of x_i)*
4. *One can show: X Hausdorff, $Y \subset X \implies$ the subspace topology is Hausdorff, and X, Y Hausdorff $\implies X \otimes Y$ Hausdorff.*

Theorem 6.50. *If X is Hausdorff then every sequence in X converges to at most one limit.*

Proof. Assume $x_1, x_2, ...$ converge to $x \in X$, and let $y \neq x$. Choose $U_x \in x, U_y \in y$ disjoint neighborhoods. Since $x_n \to n, \exists N$ such that $\forall n \geq N, x_n \in U_x$. Hence $x_n \notin U_y$ for $n \geq N$, so the sequence doesn't converge to y.

□

Remark 6.51. *There's in fact a whole hierarchy of "separation axioms": eg. a weaker one is: a topological spaces is T_1 if $\forall x \neq y \in X, \exists U_y \in y$ neighborhood such that $x \notin U_y$. Equivalently: X is $T_1 \iff \{x\}$ is closed in $X \forall x \in X$. Hausdorff $\implies T_1$, but eg. ($\mathbb{R}$, finite complement topology) is T_1 but not Hausdorff.*

Hausdorff spaces are fairly nice to work with, and ew will generally be working with this assumption. There are more subtle reasons why not ever Hausdorff topology comes from a metric, but one can give pretty good criteria for a topology to be metrizable involving further separation conditions ("normal" or T_4 + a countability condition). We'll see the Urysohn metrization theorem.

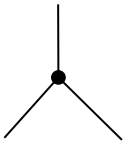
6.8 Manifolds and CW Complexes

Metric spaces are nice, but they can still be pretty nasty. (We'll see conditions such as local connectedness, local compactness, etc. come up). Algebraic topologists like to focus on even nicer spaces.

Definition 6.52. *An n-dimensional topological manifold is a topological space X such that every point $p \in X$ has a neighborhood homeomorphic to $\mathbb{R}^n$ (or equivalently, an open ball in $\mathbb{R}^n$).*

Example 6.53.

1. *$S^1 \subset \mathbb{R}^2$ is a 1d topological manifold.*
2. *Sphere and torus $\subset \mathbb{R}^3$ are 2d topological manifolds.*

Example 6.54. 

isn't a topological manifold - but it is part of a more general class of spaces called CW-complexes, built by attaching "cells" (closed balls of dimension 0, 1, ...) onto each other inductively.

We'll see more on this later when we get to algebraic topology. In decreasing order of generality,

$$\{\text{manifold}\} \subset \{\text{CW-complexes}\} \subset \{\text{metrizable}\} \subset \{\text{Hausdorff}\} \subset \{\text{topological spaces}\}.$$

6.9 Topologies on Infinite Products

Given topological spaces $X_i, i \in I$ index set. What is the natural topology on $X = \prod_{i \in I} X + i = \{(p_i)_{i \in I} | p_i \in X_i \forall i \in I\}$?

First idea:

Definition 6.55. *The **box topology** on $\prod_{i \in I} X_i$ has basis $\{\prod_{i \in I} U_i | U_i \subset K_i$ open $\forall i\}$.*

This is a basis: $\text{box} \cap \text{box} = \text{box}$, since $(\prod U_i) \cap (\prod V_i) = \prod (U_i \cap V_i)$.

This is actually too find for most purposes.

Example 6.56. *Consider the diagonal map $\Delta : \mathbb{R} \to \mathbb{R}^w = \mathbb{R}^{\mathbb{N}} (= \mathbb{R}_0 \times \mathbb{R}_1 \times ...)$ giving $\mathbb{R}^{\mathbb{N}}$ the box topology, Δ is not continuous (unlike case of finite products). Indeed, let $U = (-1, 1) \times (-\frac{1}{2}, \frac{1}{2}) \times (-\frac{1}{3}, \frac{1}{3}) \times ...$ open in box topology. $\Delta^{-1}(U) = \bigcap_{n \geq 1} (-\frac{1}{n}, \frac{1}{n}) = \{0\}$ not open in $\mathbb{R}$.*

Here's something better.

Definition 6.57. *The **product topology** on $X = \prod X_i$ has basis $\{\prod_{i \in I} U_i | U_i \subset X_i$ open, and $U_i = X_i$ for all but finitely many $i\}$.*

This is the same as the box topology if I is finite; for infinite I this is coarser. Unless otherwise specified, the product topology is the one we'll use on $\prod X_i$.

Theorem 6.58. *$f : Z \to X = \prod X_i, z \mapsto (f_i(z))_{i\in I}$ is continuous $\iff$ each component $f_i : Z \to X_i$ is continuous.*

Example 6.59. *This implies the diagonal map $\Delta : \mathbb{R} \to \mathbb{R}^{\mathbb{N}}$ is continuous, since each $\Delta_i = id$.*

Proof.

- Projection $p_i : X \to X_i$ to be the ith factor is continuous ($\forall U \subset X_i$ open, $p_i^{-1}(U)$ is open in the product topology). Hence, if f is continuous, so is $f_i = p_i \circ f$.
- Conversely, assume all f_i's are continuous, and consider basis element $\prod U_i \subset X$ where $U_i \subset X_i$, for all but finitely many i, then $f^{-1}(\prod U_i) = \{z \in Z | (f_i(z))_{i\in I} \in \prod U_i\} = \bigcap_{i\in I} f_i^{-1}(U_i)$. Each $f_i^{-1}(U) \subset Z$ is open, and all but finitely many are $= f_i^{-1}(x_i) = z$, xo can be omitted from the intersection. So $f^{-1}(\prod U_i)$ is the intersection of finitely many open set in Z, hence open.

□

Example 6.60. *Given a set X and a topological space Y, let $\mathcal{F} = \{$functions $X \to Y\} = Y^X$ with product topology. Then a sequence $f_n \in \mathcal{F}$ converges $f \in \mathcal{F}$ if and only if $\forall x \in X, f_n(x) \to f(x)$ in Y. So: the product topology is the topology of pointwise convergence.*

On products of metric spaces, there is another natural topology, finer than product but coarser than box topology - the uniform topology. This works similarly to the construction of $d_\infty(x, y) = \sup(|y_i - x_i|)$ on $\mathbb{R}^n$, but for an infinite product the sup might be infinite. So let's replace the metric on (X, d) by $\overline{d}(x, y) = \min(d(x, y), 1)$, this is still a metric and induces the same topology as of (same balls of radius ≤ 1). Now, given a metric space $(X_i, d_i)_{i\in I}$, replace each d_i by bounded metric $\overline{d_i}$, and define a metric $\overline{d_\infty}(x, y) = \sup\{\overline{d_i}(x_i, y_i) | i \in I\}$ on $\prod X_i$, which is equal to $\sup\{d_i(x_i, y_i)\}$ if it's ≤ 1, otherwise 1.

This is called the **uniform metric** and induces the **uniform topology**.

Example 6.61. *On $\mathbb{R}^\times = \{$functions $X \to \mathbb{R}\}$ (with usual distance on $\mathbb{R}$), this is $\overline{d_\infty}(f, g) = sup_{x\in X}|f(x) - g(x)|$ if ≤ 1, else 1, so $f_n \to f \iff \overline{d_\infty}(f_n, f) \to 0 \iff sup_{x\in X}|f_n(x) - f(x)| \to 0$ uniform convergence.*

Remark 6.62. *The ball of radius $r \leq 1$ around $x = (x_i)_{i\in I}$ is **contained** in $P_r(x) = \prod_{i\in I} B_r(x_i)$, but not equal to it (unless I is finite)! Indeed, $d(x_i, y_i) < r \forall i \in I$ only implies $\overline{d_\infty}(x, y) = sup_{i\in I}\{d(x_i, y_i)\} \leq r$. The ball $B_r(x)$ only contains these y for which the sup is $< r$. In fact: $B_r(x) = \bigcup_{0<r'<r} P_{r'}(x) \subset P_r(x)$... and $P_r(x)$ is not open for d_∞.*

Theorem 6.63. *The uniform topology on $\prod(X_i, d_i)$ is finer than the product topology, and coarser than the box topology (strictly if I is infinite).*

Proof.

1. Let $x = (x_i) \subset \prod X_i$ and $\prod U_i \ni x$ a basis element in the product topology, then $\forall i \exists \epsilon > 0$ such that $B_{\epsilon_i}(x_i) \subset U_i$. Without loss of generality, we can assume $\epsilon_i \leq 1 \forall i$, and $\epsilon_i = 1$ for all but finitely many i (whenever $U_i = X_i$). So $\epsilon = \inf(\epsilon_i) > 0$ and $B_\epsilon^{\overline{d_\infty}} \subset P_\epsilon(x) \subset \prod B_{\epsilon_i}(x_i) \subset \prod U_i$. So $\prod U_i$ is open in uniform topology: $\tau_{\text{product}} = \tau_{\text{uniform}}$.
2. $B_r^{\overline{d_\infty}}(x) = \bigcup_{0<r'<r} P_r(x) \implies$ balls of uniform topology are open in box topology, so $\tau_{\text{uniform}} = \tau_{\text{box}}$.

□

Remark 6.64. *On $\mathbb{R}^{\mathbb{N}}$ the product topology is actually metrizable, using a clever modification of $\overline{d_\infty}$ (see Munkres Thm. 20.5), while box isn't metrizable (Munkres section 21). On uncountable products, neither box nor product are metrizable.*

The notion of uniform convergence is important in real analysis because it is well behaved with respect to continuity and differentiability. For example:

Theorem 6.65. *Given a topological space X, metric space Y, and a sequence of functions $f_n : X \to Y$, if f_n is continuous $\forall n$ and $f_n \to f$ uniformly then f is continuous.*

Proof. Let $V \subset Y$ open, $p \in f^{-1}(V)$. $\exists \epsilon > 0$ such that $B_\epsilon(f(x)) \subset V$. Let N be such that $\sup_{q \in X} d(f_N(q), f(q)) < \frac{\epsilon}{3}$. Let $U \ni p$ open such that $q \in U \implies d(f_N(p), f_N(q)) < \frac{\epsilon}{3}$ (continuity of f_N). Then using triangle inequality: $\forall q \in U$,

$$d(f(p), f(q)) \leq d(f(p), f_N(p)) + d(f_N(p), f_N(q)) + d(f_N(q), f(q)) < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon.$$

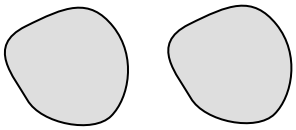
So $U \subset f^{-1}(B_\epsilon(f(p))) \subset f^{-1}(V)$.

□

Corollary 6.66. *$\{$continuous $f : X \to Y\}$ is a closed subspace of $(\mathcal{F}(X, Y) = Y^X$, uniform topology).*

6.10 Connected Spaces

Definition 6.67. *A topological space X is **connected** if it cannot be written as $X = U \cup V$ where U, V are disjoint nonempty open sets. (such a decomposition is called a **separation** of X).*



not connected

Proposition 6.68. $[0,1] \subset \mathbb{R}$ *(standard topology is connected).*

Proof. Assume $[0,1] = U \cup V$ separation. Without loss of generality, $u \in U$. Let $a = \sup\{x \in [0,1] \text{ such that } [0,x) \subset U\}.0 \in U, U$ open $\implies [0,\epsilon) \subset U$ for some $\epsilon > 0$, so $a > 0$.

We cant have $a \in V$, since V is open this would imply $(a-\epsilon, a] \subset V$ for some $\epsilon n > 0$, hence $[0,x)$ is not a subset of U for $x?a - \epsilon$, hence $\sup\{x \text{ such that } [0,x) \subset U\} \leq a - \epsilon$, contradiction. So $a \in U$. But if $a < 1$, U open, $U \ni a \implies \exists \epsilon > 0$ such that $(a-\epsilon, a+\epsilon) \subset U$, and by definition of $a, \exists x > a-\epsilon$ such that $[0,x) \subset U$. Hence $[0, a+\epsilon) \subset U$, contradicting definition of a.

Hence $a = 1$ and since U is open, $\exists \epsilon > 0$ such that $(1-\epsilon, 1] \subset U$, and by definition of a, $\exists x > 1 - \epsilon$ such that $[0,x) \subset U$, hence $U = [0,1]$ and $V = \emptyset$, contradiction.

□

Example 6.69. $[0,1) \cup (1,2]$ *is not connected, since* $[0,1)$ *and* $(1,2]$ *are open in subspace topology and provide a separation. More generally,* $x < y < z \in \mathbb{R}$, $x, z \in A, y \notin A \implies A$ *disconnected.*

Theorem 6.70. $f : X \to Y$ *continuous,* X *connected* $\implies f(X) \subset Y$ *is connected.*

Proof. If $U \cup V$ is a separation of $f(X)$, then $f^{-1}(U) \cup f^{-1}(V)$ is a separation of X, contradiction. (subspace topology: $U = f(X) \cap U' = \emptyset, U'$ open in $Y \implies f^{-1}(U) = f^{-1}(U') \neq \emptyset$ open in X; $f^{-1}(U) \cap f^{-1}(V) = f^{-1}(U \cap V) = \emptyset$).

□

A corollary:

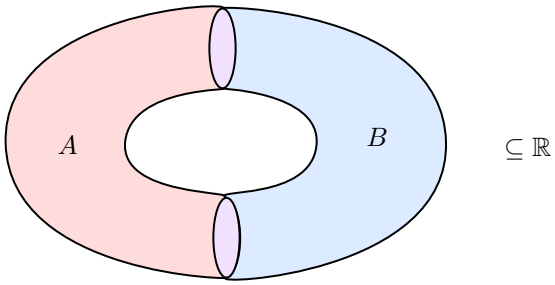
Theorem 6.71 (Intermediate Value Theorem)**.** *Let* X *be a topological space with a function* $f : X \to \mathbb{R}$ *continuous. If* $a, b \in X$ *and* r *lies between* $f(a)$ *and* $f(b)$*, then* $\exists c \in X$ *such that* $f(c) = r$.

Proof. Since X is connected, so is $f(X)$. If $r \notin f(X)$ then $U = (-\infty, r) \cap f(X)$ and $V = (r, \infty| \bigcap f(X)$ gives separation of $f(X))$ (one contains $f(a)$ and the other contains $f(b)$) - contradiction. So $r \in f(X)$.

□

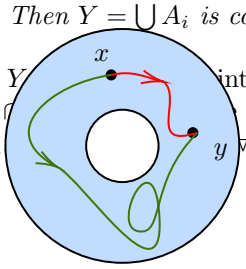
Proposition 6.72. *$A, B \subset X$ connected (for subspace topology) does not imply $A \cap B$ connected.*

Example 6.73. *Take two halves of a torus and stick them together into a donut.*



But things are better for unions of connected sets, provided they overlap.

Theorem 6.74. *$A_i \subset X$ connected subspaces, all containing some point $p \in X$ (ie. $\bigcap A_i \neq \emptyset$). Then $Y = \bigcup A_i$ is connected.*



Proof. Assume Y [illegible]int, open in Y. Without loss of generality, $p \in U$. Then $U \cap$ [illegible] disjoint, open in A_i. Since A_i is connected and $p \in U \cap A_i$ [illegible] $\forall i$. Hence $Y = \bigcap A_i \subset U$ (and $V = \emptyset$). So Y is connected.

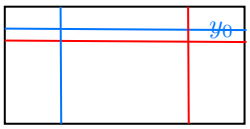
two paths

□

Corollary 6.75. *$\mathbb{R}$ is connected; so are open, half-open, and closed intervals in $\mathbb{R}$.*

Theorem 6.76. *X, Y connected $\implies$ $X \times Y$ is connected.*

Proof. Fix $(x_0, y_0) \in X \times Y$. Then $\forall x \in X, A_x = (x \times \{y_0\}) \cup (\{x\} \times Y)$ is connected by previous theorem (both pieces contain (x, y_0)) and now $X \times Y = \bigcup_{x \in X} A_x$ (all containing (x_0, y_0)) $\implies$ $X \times Y$ is connected.



□

In fact, more is true:

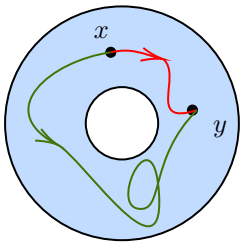
Theorem 6.77. *$X_i, i \in I$ connected $\implies$ $\prod_{i\in I} X_i$ with product topology is connected.*

This is false for uniform and box topologies: eg. $\mathbb{R}^I = \{$ functions $I \to \mathbb{R}\}$ for infinite I. Say $f : I \to \mathbb{R}$ is bounded if $f(I) \subset \mathbb{R}$ bounded subset. Then $\{\text{bounded}\} \cup \{\text{unbounded}\}$ is a separation of $\mathbb{R}^I$ in uniform topology.

6.11 Path-connectedness

Definition 6.78. *X topological space, $x, y \in X$, a **path** from x to y is a continuous map $f : [a, b] \to X$ such that $f(a) = x$ and $f(b) = y$.*

Definition 6.79. *X is **path-connected** if every pair of points in X can be joined by a path.*



two paths $x \to y$

Note: The relation $x \sim y \implies x$ and y can be connected by a path is an equivalence relation:

1. $x \sim x$ (constant path $f(t) = x$).
2. $x \sim y \iff y \sim x$ (backwards path $f(-t)$)
3. $x \sim y$ and $y \sim z \implies x \sim z$ (concatenate paths)

The equivalence classes are called the **path components** of X - we will return to this in algebraic topology.

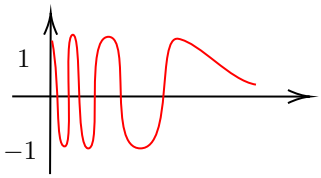
Theorem 6.80. *If X is path connected then X is connected.*

Proof. Assume not, ie. $X = U \sqcup V$ disjoint open, $x \in U, y \in V$. Pick a path $f : [a, b] \to X$ connecting x to y. Then $[a, b] = f^{-1}(U) \sqcup f^{-1}(V)$ open subsets. This contradicts the connectedness of $[a, b]$.

□

The converse is false in general, but true for nice enough spaces (eg. CW-complexes).

Example 6.81. *The "topologist's sine curve": let* $S = \{(x, y) | y = \sin\left(\frac{1}{x}\right), x > 0\} \cup \{(0,0)\} \subset \mathbb{R}^2$. *The "main" point of* S_0 *is connected, since it's the image of* $\mathbb{R}_+$ *(connected) under the continuous map* $x \mapsto \left(x, \frac{1}{x}\right)$.



Hence, S *is connected: if* $S = U \cup V$ *disjoint open, then* $S_0 = (U \cap S_0) \cup (V \cap S_0)$ *disjoint and open* $\implies$ *one of them (eg.* $V \cap S_0$*) is empty.* $V \subset S - S_0 = \{(0,0)\}$. *But* $\{(0,0)\}$ *is not open in* S*, so in fact* $V = \emptyset$.

On the other hand, S *is not path connected: there's no path connecting* $\left(\frac{1}{\pi}, 0\right)$ *to* $(0,0)$. *(We can prove this later using compactness: the image of such a path is a closed subset of* $\mathbb{R}^2$*, but* S *isn't:* $(0,1)$ *is a limit point of* S *not in* S*).*

However, for nice enough spaces the two notions are equivalent.

Theorem 6.82. $A \subset \mathbb{R}^n$ *open* $\implies$ A *is connected if and only if* A *is path connected.*

Proof. Already seen: path connected $\implies$ connected. We show: not path connected $\implies$ not connected.

Assume A open in $\mathbb{R}^n$: then the path components of A are open. Indeed, if $x \in A$ then $\exists r \in 0$ such that $B_r(x) \subset A$, and any two points of $B_r(x)$ can be connected inside A by a straight line segment. So all of $B_r(x)$ is in the same path component.

Now: if A is not path connected then $A =$ (one path connected)$\cup$($\bigcup$ all other path components) gives a separation.

□

This implies similar results for other classes of spaces, eg. topological manifolds and CW-complexes.

For these kinds of spaces, path-components are also connected components, ie. they give a partition of X into disjoint connected open (and closed) subsets. Such a partition only exists if X is "locally connected" ie. the topology has a basis consisting of connected open subsets. Counterexample: $\mathbb{Q} \subset \mathbb{R}$ isn't locally connected. (each point of Q is its own path component, but these aren't open).

6.12 Compactness

Compactness is a "finiteness/boundedness" property of nice topological spaces such as closed bounded intervals $[a, b] \subset \mathbb{R}$, or more generally, closed bounded

subsets of $\mathbb{R}^n$. Any continuous map $f : K \to \mathbb{R}$ (where K is compact) achieves its maximum and minimum. The definition isn't very intuitive.

Definition 6.83. *An **open cover** of a topological space X is a collection of open subsets $(U_i)_{i\in I}$ such that $\bigcup_{i\in I} U_i = X$.*

Definition 6.84. *X is compact if every open cover $(U_i)_{i\in I}$ of X admits a finite subcover, ie. $\exists i_1, ..., i_n$ such that $X = U_{i_1} \cup ... \cup U_{i_n}$.*

Showing a space is not compact is much easier than showing it is.

Example 6.85. *$\mathbb{R}$ is not compact: the open cover $\mathbb{R} = \bigcup_{n\in\mathbb{N}}(n-1, n+1)$ has no finite subcover. Neither is $(0, 1]$ with subspace topology $(0, 1] = \bigcup_{n\in\mathbb{N}}(\frac{1}{n}, 1]$ has no finite subcover.*

Example 6.86. *$X = \{0\} \cup \{\frac{1}{n}, n \in \mathbb{Z}_+\}$ is compact: given any open cover $X = \bigcup_{i\in I} U_i$, let i_0 be such that $0 \in U_{i_0}$, then U_{i_0} also contains $\frac{1}{n}$ for large $n \geq N$, hence $U_{i_0}, ..., U_{i_N}$ containing $1, \frac{1}{2}, ..., \frac{1}{N}$ and U_{i_0} gives a finite subcover.*

Theorem 6.87. *If X is compact and $f : X \to Y$ is continuous, then $f(X) \subset Y$ is compact.*

Proof. Let $\bigcup_{i\in I} U_i$ open cover of $f(X)$. Then $\bigcup_{i\in I} f^{-1}(U_i)$ is an open cover of X, hence $\exists i_1...i_n$ such that $f^{-1}(U_{i_1}) \cup ... \cup f^{-1}(U_{i_n}) = X$. So $\forall x \in X$, $f(x) \in U_{i_1} \cup ... \cup U_{i_n}$, ie. $f(X) \subset U_{i_1} \cup .. \cup U_{i_n}$ finite subcover.

□

Remark 6.88. *An open cover of $f(X) \subset Y$ with subspace topology $\iff U_i \subset Y$ open, $f(X) \subset \bigcup_{i\in I} U_i$.*

Once we know subsets of $\mathbb{R}^n$ are compact if and only if closed and bounded, taking $Y = \mathbb{R}$, this gives the extreme value theorem. To get started on this right away.

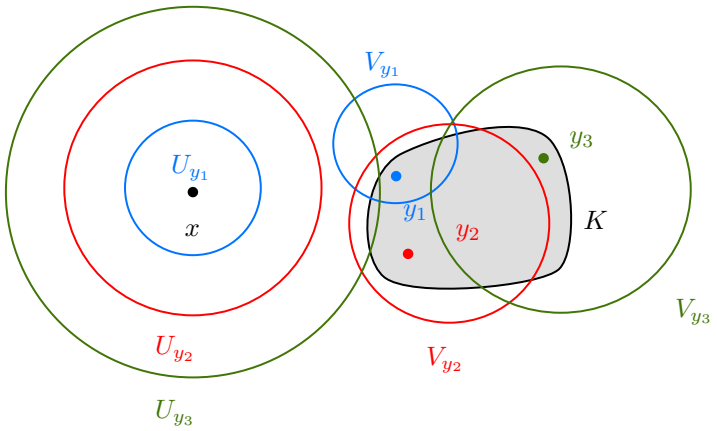
Theorem 6.89. *$[0, 1]$ (with subspace topology $\subset \mathbb{R}$) is compact.*

Proof. Let $\{U_i\}_{i\in I}$ open cover of $[0, 1]$. Let $A = \{x \in [0, 1] | \exists$ finite subcover $U_{i_1} \cup ... \cup U_{i_n} \supset [0, x]\}$. $A \neq \emptyset$ (contains 0). We want to show $1 \in A$. Let $a = \sup(A) \in [0, 1]$.

First we show $a \in A$: $\exists i_0$ such that $a \in U_{i_0}$; since U_{i_0} is open, $\exists \epsilon > 0$ such that $B_\epsilon(a) \subset U_{i_0}$. On the other hand, $a = \sup A$, so $\exists x \in A$ such that $x > a - \epsilon$, and a finite subcover $[0, x] \subset U_{i_1} \cup ... \cup U_{i_n}$. Therefore $[0, a] \subset U_{i_1} \cup ... \cup U_{i_n} \cup U_{i_0}$, and $a \in A$.

Next, assume $a < 1$: since $a \in A, \exists i_1, ..., i_n$ such that $[0, a] \subset U_{i_1} \cup ... \cup U_{i_n}$, which is open, so $\exists \epsilon > 0$ such that $B_\epsilon(a) \subset U_{i_1} \cup ... \cup U_{i_n}$, hence $U_{i_1} \cup ... \cup U_{i_n}$ covers $[0, x]$ for some $x > a$ (eg. $x = a + \frac{\epsilon}{2}$ if ≤ 1, else 1), contradicts $\sup(A) = a$.

So $a = 1 \in A, \exists$ finite subcover.



□

Theorem 6.90. *X compact, $F \subset X$ closed $\implies$ F is compact.*

Proof. Given an open cover of F, ie. $U_i \subset X$ open, $\bigcup_{i \in I} U_i \supset F$, let $V = X \backslash F$ open: then $\{U_i, i \in I\} \cup \{V\}$ is an open cover of X, hence $\exists$ finite subcover. Discarding V, this gives a finite subcover for F.

□

The converse is true in Hausdorff spaces!

Theorem 6.91. *X Hausdorff, $K \subset X$ compact $\implies$ K is closed in X.*

Proof. We show that $X \backslash K$ is open. Let $x \in X \backslash K$. Since X is Hausdorff, $\forall y \in k \exists U_y \ni x, V_y \ni y$ disjoint open subsets. Now $K \subset \bigcup_{y \in K} V_y$ is an open cover, so by compactness $\exists y_1, ..., y_n$ such that $K \subset V_{y_1} \cup ... \cup V_{y_n}$. Let $U = U_{y_1} \cup ... \cup U_{y_n} \ni x$ open. Then $U \cap (V_{y_1} \cup ... \cup V_{y_n}) = \emptyset$, so $U \cap K = \emptyset$. Hence $\forall x \in X \backslash K, \exists U$ open $\ni x$ such that $U \subset X \backslash K$.

□

Remark 6.92. *We've actually shown more: X Hausdorff, $K \subset X$ compact, $x \in X \backslash K \implies \exists$ disjoint open subsets $U \ni x, V \supset K, U \cup V = \emptyset$. Ie. can separate points from compact subsets.*

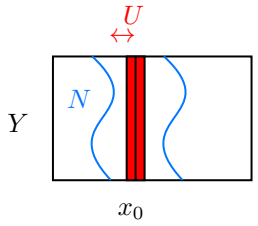
If we tried this for an arbitrary subset of X, we'd find that $\bigcup_{y \in K} U_y$ isn't a neighborhood of x anymore. Compactness lets us reduce on infinite process to a finite one.

Example 6.93. *When X isn't Hausdorff, $K \subset X$ compact does not imply K closed in X: eg. $X = \mathbb{R}$ with finite complement topology: any subset $K \subset X$ is compact. Indeed, a nonempty open subset contains all but finitely many points,*

so given an open cover it is easy to find a finite subcover: take one nonempty U_i, with finite complement $\{p_1 ... p_k\}$, then take U_{i_j} containing p_j for $j = 1, ..., k$.

Another instance of compactness allowing us to intersect infinitely many opens (or rather reduce to a finite intersection) is the tube lemma.

Proposition 6.94 (The Tube Lemma)**.** *Let X topological space, Y compact topological space, $x_0 \in X$, if $N \subset X \times Y$ is open and $\{x_0\} \times Y \subset N$, then there exists a neighborhood U of x_0 in X such that $U \times Y \subset N$.*



Proof. $\forall y \in Y, (x_0, y) \in N$ open, so $\exists$ basis open $U_y \times V_y$ (U_y neighborhood of x_0 in X, V_y neighborhood of y in Y) such that $(x_0, y) \in U_y \times V_y \subset N$.

Now: $\bigcup_{y \in Y} V_y = Y$ open cover (Rmk: $\left(\bigcap_{y \in Y} U_y\right) \times Y \subset N$, but $\bigcup_{y \in Y} U_y$ not open!). Since Y is compact, $\exists y_1, ..., y_n \in Y$ such that $Y = V_{y_1} \cup ... \cup V_{y_n}$. Let $U = U_{y_1} \cap ... \cap U_{y_n}$. Then U is a neighborhood of x_0 in X, and $U \times Y = \bigcup_{i=1}^{n} U_{y_1} \times V_{y_i} \subset N$.

□

Theorem 6.95. *X, Y compact $\implies$ $X \times Y$ is compact.*

Proof. Let $\{A_\alpha\}$ be an open cover of $X \times Y$. For any given $x \in X, \{x\} \times Y$ is compact so $\exists$ finite subcollection $A_{x,1}, ..., A_{x,n(x)}$ which suffices to cover $\{x\}$ or Y. $A_{x,1} \cup ... \cup A_{x,n(x)}$ is open, so by the tube lemma $\exists U_x \ni x$ neighborhood in X such that $A_{x,1} \cup ... \cup A_{x,n(x)} \supset U_x \times Y$. Now X is compact, and $\{U_x\}_{x \in X}$ form an open cover, so $\exists x_1, ..., x_k \in X$ such that $X = U_{x_1} \cup ... \cup U_{x_k}$. Now $A_{x_i,j}, 1 \leq i \leq k, 1 \leq j \leq n(x_i)$ is a finite subcover for $X \times Y$.

□

Theorem 6.96. *$K \subset \mathbb{R}^n$ is compact if and only if K is closed and bounded.*

Proof.

- If $K \subset \mathbb{R}^n$ is compact then it is closed (by above thm: $\mathbb{R}^n$ Hausdorff) and bounded: $K \subset \bigcup_{r>0} B_r(0)$ open cover $\implies$ $\exists$ finite subcover $\implies$ $\exists R > 0$ such that $K \subset B_R(0)$.

- If $K \subset \mathbb{R}^n$ is closed and bounded, then it's a closed subset of $[-R, R]^n$ for some $R > 0$. $[-R, R]^n$ is a finite product of compact sets ($[R, -R] \simeq [0, 1]$) hence compact; a closed subset of a compact is compact.

□

Remark 6.97. *Closed and bounded are necessary conditions for compactness of a subspace of any metric space but in "most" metric spaces, closed + bounded does not imply compact. There are easy counterexamples.*

More interesting: let V be any infinite-dimensional vector space with a norm, $d(v, v') = ||v - v'||$. Eg. $\mathcal{F} = C^0([a, b], \mathbb{R})$ continuous functions with sup norm $d(f, g) = \sup|f - g|$ (uniform topology). Then $\overline{B} = \{v \in V | \, ||v|| \leq 1\}$ is closed and bounded but never compact (proof uses sequential compactness).

We now look at applications of compactness. We saw earlier that

Theorem 6.98. *If X is compact and $f : X \to Y$ is continuous, then $f(X) \subset Y$ is compact.*

Corollary 6.99 (Extreme Value Theorem)**.** *X compact (nonempty), $f : X \to \mathbb{R}$ continuous $\implies$ f attains its maximum and minimum.*

Example 6.100. *(X, d) metric space, $A \subset X$ nonempty, $x \in X \implies$ define $d(x, A) = inf\{d(x, a)|a \in A\} \geq 0$. If A is compact then the inf is always achieved! (see Munkres 27.2). Similarly, the diameter of a bounded subset, $diam(A) = sup\{d(x, y)|x, y \in A\}$. The sup is attained for A compact ($d : A \times A \to \mathbb{R}$ continuous, achieve its max).*

Another corollary:

Corollary 6.101. *If X is compact and Y is Hausdorff, then any continuous bijection $f : X \to Y$ is a homeomorphism.*

Proof. We need to check f^{-1} is continuous as well (so $U \subset X$ open $\iff f(U) \subset Y$ open) $U \subset X$ open $\implies X \backslash U$ closed hence compact $\implies f(X \backslash U) = Y \backslash f(U)$ compact. Since Y is Hausdorff this implies $Y \backslash f(U)$ is closed, ie. $f(U)$ open in Y. (We've seen that with such assumptions a continuous bijection need not be a homeomorphism, eg. $[0, 2\pi) \to S^1, t \mapsto (\cos t, \sin t)$).

□

In metric spaces, compactness implies uniform estimates.

Lemma 6.102 (Lebesgue Number Lemma)**.** *(X, d) compact metric space $(U_i)_{i \in I}$ open cover of $X \implies \exists \delta > 0$ such that any subset of diameter $< \delta$ is entirely contained in a single open U_i.*

Proof. By compactness, can assume (U_i) is a finite cover $U_1 \cup ... \cup U_n$. The function $f(x) = \frac{1}{n} \sum_{i=1}^{n} d(x, X \backslash U_i)$ is continuous so achieves its min, which is therefore > 0 ($\forall x \in X \exists i$ such that $x \in U_i$ and then $d(x, X \backslash U_i) > 0$). Hence

$\exists \delta > 0$ such that $f(x) \geq \delta \forall x \in X$. Thus $\forall x \in X \exists U_i$ such that $d(x, X \backslash U_i) \geq \delta$, ie. $B_\delta(x) \subset U_i$. Since a subset of diameter $< \delta$ is contained in a ball of radius δ, the result follows.

□

This is the magic of compactness!

Counterexamples: $\mathbb{R} = \bigcup$ intervals with overlaps of lengths $\to 0$, eg. $\bigcup_{n \in \mathbb{Z}} (n - 1, n + 1 + \epsilon_n)$ with $\epsilon_n \to 0$. For example, $xy = 1$ and $y = 0$ has distance $\to 0$ as $x \to \infty$.

THis only makes sense for metric spaces! No notion of uniform size of neighborhood without a metric.

Definition 6.103. *$f : (X, d_X) \to (Y, d_Y)$ is **uniformly continuous** if $\forall \epsilon > 0, \exists \delta > 0$ such that $\forall p, q \in X, d_X(p, q) < \epsilon \implies d_y(f(p), f(q)) < \epsilon$.*

Compare with continuity: the same δ must work for every p!

Theorem 6.104. *If X and Y are metric spaces, $f : X \to Y$ continuous, and X is compact, then f is uniformly continuous.*

Proof. Take $\epsilon > 0$ and consider open cover of Y by balls of radius $\frac{\epsilon}{2}$. (so if $f(p), f(q)$ land in same ball, they're less than ϵ apart). $X = \bigcup_{y \in Y} f^{-1}(B_{\frac{\epsilon}{2}}(y))$ open cover, so by Lebesgue number lemma $\exists \delta > 0$ such that if $d_X(p, q) < \delta$ then they lie in the same element of the cover, hence $d_Y(f(p), f(q)) < \epsilon$.

□

6.13 Alternative Notions of Compactness

Definition 6.105.

- *X is a **limit point compact** if every infinite subset of X has a limit point.*
- *X is **sequentially compact** if every sequence $\{p_n\}$ in X has a convergent subsequence.*

Example 6.106. *In $\mathbb{R}$, $\{\frac{1}{n}, n \geq 1\} \cup \mathbb{Z}_+$ has a limit point (0) and the sequence $1, 2, \frac{1}{2}, 3, \frac{1}{3}, ...$ has a convergent subsequence $\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{4}\right)$ so does $0, 1, 0, 1, 0, 1, ...$ (eg. subsequence $0, 0, ...$). But $\mathbb{Z} \subset \mathbb{R}$ has no limit point and the sequence $1, 2, 3, ...$ doesn't have a convergent subsequence, so $\mathbb{R}$ is neither limit point compact nor sequentially compact.*

Theorem 6.107. *X is compact $\implies$ X is limit point compact.*

Proof. Assume X is not limit point compact, ie. $\exists A \subset X$ infinite with no limit point. Since A contains all of its limit points (there are none), A is closed in X, hence compact. However, $\forall a \in A$, a isn't a limit point so $\exists U_a \ni a$ neighborhood

of a such that $U_a \cap A = \{a\}$. $(U_a)_{a\in A}$ is now an infinite open cover of A, without any finite subcover since each $a \in A$ only belongs to U_a and not to any other element of the cover, contradiction.

□

Theorem 6.108. *X sequentially compact $\implies$ X limit point compact.*

Proof. Given $A \subset X$ infinite subset, pick a sequence of distinct points of A and take a convergent subsequence $\implies \exists\{a_n\}$ sequence in A, $a_n \neq a_m \forall n \neq m$, converging to some limit $a \in X$. Then every neighborhood of a contains a_n for all large n, hence only many points of A, including some $\neq a$. So a is a limit point of A.

□

The converse implications don't hold in general, but in metric spaces all three notions coincide! (and hence also for subspaces of metric spaces...)

Theorem 6.109. *For a metric space (X, d), X compact $\implies$ X limit point compact $\iff$ X sequentially compact.*

Proof.

- Compact $\implies$ limit point compact already done
- Limit point compact $\implies$ sequentially compact: suppose X metric space and limit point compact, and consider a sequence $x_1, x_2, ...$ in X. If $\{x_1, x_2, ...\}$ finite, then $\exists x \in X$ such that $x_n = x$ for infinitely many n, which gives a subsequence that converges to x. Otherwise $\{x_1, x_2, ...\}$ is infinite, so has a limit point a. So: $\forall r > 0 \exists n$ such that $0 < d(a, x_n) < r$. First choose $n_1 \in \mathbb{N}$ such that $x_{n_1} \in B_1(a)$, then inductively given $n_1, .., n_{k-1}$, let $\sigma_k = \min\{d(x_i, a) | i \leq n_{k-1} \text{ and } x_i \neq a\} > 0$ and $r_k = \min\left(\frac{1}{k}, \delta_k\right)$. Then take n_k such that $0 < d(a, x_{n_k}) < r_k$. By construction: $n_k > n_{k-1}$ and $d(a, x_{n_k}) < \frac{1}{k} \implies x_{n_1}, x_{n_2}, ...$ is a subsequence converging to a.
- Sequentially compact $\implies$ compact; this is the hardest part. First we show:

Claim: If X metric space is sequentially compact, then $\forall \epsilon > 0$ X can be covered by finitely many open balls of radius ϵ.

(as we expect if X is to be compact: $X = \bigcup_{x\in X} B_\epsilon(x)$ should have a finite subcover!)

Claim Proof. Assume not, and choose $x_1 \in X$, then inductively choose $x_n \in X \setminus \bigcup_{i=1}^{n-1} B_\epsilon(x_i)$ (if this isn't possible then we've covered X by finitely many balls). This yields a sequence in X, which by sequential compactness must have

a convergent subsequence. But this is impossible since no two terms of the two sequence are within ϵ of each other, contradiction.

Claim: If X metric space is sequentially compact then every open cover has a Lebesgue number ($\exists \epsilon > 0$ such that any subset of diameter $< \delta$ is entirely in one U_i).

(we've seen this hold for compact metric spaces, so it should hold!)

Claim Proof. Suppose $\exists$ open cover $(U_i)_{i\in I}$ with no Lebesgue number, ie. $\forall n \geq 1, \exists C_n \subset X$ with diameter $< \frac{1}{n}$ which isn't contained in any single U_i. Take $x_n \in C_n$. By sequential compactness, $\exists$ subsequence (x_{n_k}) of (x_n) that converges to some $a \in X$. Now $a \in U_{i_0}$ for some $i_0 \in I$ and so $\exists \epsilon > 0$ such that $B_\epsilon(a) \subset U_{i_0}$. Take k sufficiently large so that $\frac{1}{n_k} < \frac{\epsilon}{2}$ and $d(x_{n_k}, a) < \frac{\epsilon}{2}$. Since c_{n_k} has diameter $< \frac{\epsilon}{2}, c_{n_k} \subset B_{\frac{\epsilon}{2}}(x_{n_k}) \subset B_\epsilon(a) \subset U_{i_0}$, contradiction.

This proof illustrates how arguments using sequential compactness are often more intuitive than those involving open covers: "if some property fails to hold uniformly, take a sequence of points where things get worse and worse, extract a convergent subsequence, and see what goes wrong at the limit."

Now we can finish proving sequentially compact $\implies$ compact: Given an open cover $X = \bigcup_{i\in I} U_i$, by lemma 2, $\exists \epsilon > 0$ such that every subset of diameter $< \delta$ is entirely inside a single U_i. Fix $\epsilon \in \left(0, \frac{\delta}{2}\right)$: by lemma 1 X is covered by finitely many ϵ-balls. Each of these has diameter $\leq 2\epsilon < \delta$, so is contained in some U_i. This gives a finite subcover, replacing each ϵ-ball by one U_i containing it (and discarding the rest of the U_i's).

□

Theorem 6.110. *Every compact metric space (X, d) is complete, ie. every Cauchy sequence converges.*

Proof. Let (x_n) Cauchy sequence, by sequential compactness $\exists$ subsequence $x_{n_k} \to x \in X$. Now $\forall \epsilon > 0 \exists N$ such that $\forall m, n \geq N, d(x_m, x_n) < \frac{\epsilon}{2}$. $\exists n_k \geq N$ such that $d(x_{n_k}, x) < \frac{\epsilon}{2}$. Hence $\forall n \geq N, d(x_n, x) \leq d(x_n, x_{n_k}) + \dim(x_{n_k}, x) < \epsilon$.

□

Corollary 6.111. $\mathbb{R}, \mathbb{R}^n$ *(with usual distances) are complete.*

Proof. Every Cauchy sequence is bounded, hence contained in a compact subset, hence convergent.

□

Corollary 6.112. $\mathbb{R}^X = \{$*functions* $X \to \mathbb{R}\}$ *with uniform metric is complete.*

Proof. Given a Cauchy sequence $\{f_n\}$ (ie. $\forall \epsilon > 0 \exists N$ such that $m, n \geq N \implies \sup|f_n - f_m| < \epsilon$). $\forall x \in X, \{f_n(X)\}$ is a Cauchy sequence in $\mathbb{R}$. ($|f_n(x) - f_m(x)| \leq \sup|f_n - f_m| < \epsilon$) hence converges to some limit $f(x)$ (ie. we have a pointwise limit). Now: given $\epsilon > 0$, take N such that $m, n \geq N \implies \sup_x|f_n(x) - f_m(x)| < \epsilon$. Then $\forall n \geq N, \forall x \in X, |f_n(x) - f(x)| = \lim_{m\to\infty}|f_n(x) - f_m(x)| \leq \epsilon$, ie. $\forall n \geq N, \sup|f_n - f| \leq \epsilon$ which implies $f_n \to f$ uniformly.

□

When X is a topological space, we've seen that uniform limits of continuous functions are continuous, so we also have completeness of $C^0(X, \mathbb{R}) = \{\text{continuous functions}\} \subset \mathbb{R}^X$, uniform topology. More generally: closed subsets of complete metric spaces are complete!

6.14 Compactification

Definition 6.113. *A **compactification** of a (Hausdorff) topological space X is a compact (Hausdorff) space Y with an inclusion $i : X \hookrightarrow Y$ which is an embedding (ie. homeomorphism onto its image, ie. topology on $X \equiv$ subspace topology of $i(X) \subset Y$), with X open and dense in Y ($\overline{X} = Y$).*

Example 6.114. *$\mathbb{R}^n \to \mathbb{R}^n \cup \{\infty\}$ as in HW2; this is in fact homeomorphic to S^n. This is not the only option: eg. $(0, 1) \simeq \mathbb{R}$ compactifies to $[0, 1]$ of S^1.*

$(0, 1) \times (0, 1) \simeq \mathbb{R}^2$: eg. $[0, 1] \times [0, 1], S^2$, torus $\simeq S^1 \times S^1$.

The **one-point compactification**, if exists, is unique. Let $Y = X \cup \{\infty\}$ (add a new point). The requirements of a compactification imply:

- a subset $U \subset X$ is open in Y if and only if it is open in X (subspace topology $\simeq \tau_X$).
- a subset V containing ∞ is open in Y if and only if $Y \backslash V$ is closed, hence compact (we want Y compact), and a subset of X (since $\infty \in V$).

Definition 6.115.

$$\tau_Y = \{U \subset X \text{ open}\} \cup \{Y \backslash K | K \subset X \text{ compact}\}$$

Theorem 6.116. *τ_Y is a topology on $Y = X \cup \{\infty\}$, and Y is a compactification of X (in particular, Y is compact).*

Proof.

- Axioms of topology: case by case for U's and $(Y - K)$'s. Arbitrary unions and finite intersections of a single type of open are still of the same type (note: $\bigcap(Y - K_i) = Y - (\bigcap K_i)$, a finite union of compact subsets of X is compact). Moreover, $U \cap (Y - K) = U \cap (X - K)$ open $\subset X$, $U \cup (Y - K) = Y - (K \cap (X - U))$ closed in K hence compact.

- Y is compact: if $(A_i)_{i\in I}$ open cover of Y, then $\infty \in A_{i_0} = Y - K$ for some $i_0 \in I$, and now the $(A_i \cap K)$ form an open cover of $K \implies \exists i_1, ..., i_n$ such that $A_{i_1} \cup ... \cup A_{i_n} \supset K$. Thus $Y = A_{i_0} \cup (A_{i_1} \cup ... \cup A_{i_n})$ finite subcover.

□

However, this Y is not always Hausdorff! One-point compactifications are only useful if Hausdorff.

Definition 6.117. *X is **locally compact** if $\forall x \in X, \exists K$ compact $\subset X$ which contains a neighborhood of x.*

Example 6.118.

- *$\mathbb{R}$ is locally compact ($x \in \mathbb{R} \implies x \in int([x-1, x+1])$), so is $\mathbb{R}^n$. $\mathbb{R}^\infty$ isn't (for any of usual topologies). Neither is $\mathbb{Q}$ with usual topology ($\subset \mathbb{R}$).*

Theorem 6.119. *The one-point compactification $Y = X \cup \{\infty\}$ is Hausdorff if and only if X is locally compact and Hausdorff.*

Proof.

- X Hausdorff $\iff$ we can separate points of $X \subset Y$ by open subsets (in X or in Y).
- X locally compact $\iff \forall x \in X \exists$ open $U \ni x, Y - K \ni \infty$ such that $U \subset K$ ie. $U \cap (Y - K) = \iff$ we can separate point of X from ∞ by open subsets in Y.

□

6.15 Countability Axioms

Definition 6.120. *X is **first-countable** if $\forall x \in X, \exists$ countable basis of neighborhoods at x, ie. $\exists U_1, U_2, ...$ open $\ni x$ such that every neighborhood $V \ni x$ contains one of the U_n.*

Example 6.121. *Metric spaces are first-countable: at $x \in X$ take $U_n = B_{\frac{1}{n}}(x)$.*

In a first-countable space, $x \in \overline{A} \iff \exists$ sequence $x_n \in A, x_n \to x$ (else only $\Leftarrow$).

Definition 6.122. *X is **second-countable** if its topology has a countable basis.*

Example 6.123.

- *$\mathbb{R}^n$ is second-countable, eg. basis $\{B_r(x), x \in \mathbb{Q}^n, r \in \mathbb{Q}_+\}$ or $\{\prod(a_i, b_i) | a_i, b_i \in \mathbb{Q}\}$. $\mathbb{R}^\omega$ product topology is second-countable (basis = products of finite number of (a_i, b_i), $a_i < b_i \in \mathbb{Q}$ and all remaining factors are $\mathbb{R}$), while uniform topology isn't (because $\exists$ uncountable many disjoint open subsets: balls of radius $\frac{1}{2}$ centered at $\{0,1\}^\omega$).*

Second-countable $\implies$ $\exists$ countable dense subset (eg. take one point in each basis open!). The converse holds for metric spaces (take balls of radius $\frac{1}{n}$ around points of the sense subset) but is false in general ($\mathbb{R}_\ell$ is first-countable, has countable dense subset, but $\nexists$ countable basis).

6.16 Regular and Normal Spaces

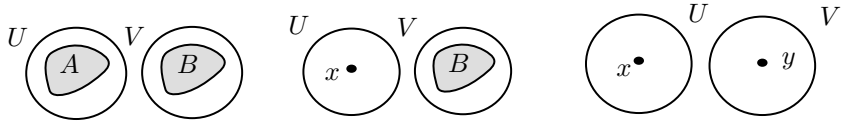
Now, we introduce some stronger separation axioms.

Definition 6.124.

Definition 6.125. *Suppose that one-point subsets $\{x\} \subset X$ are closed (i.e., T_1 space). We then say:*

- *X is **regular** if for every point $x \in X$ and every closed set $B \subset X$ that is disjoint from x, there exist disjoint open sets U containing x and V containing B, i.e., $x \in U$ and $B \subset V$.*
- *X is **normal** if for any two disjoint closed sets $A, B \subset X$, there exist disjoint open sets U containing A and V containing B, i.e., $A \subset U$ and $B \subset V$.*

Metrizable $\implies$ Normal (T_4) $\implies$ Regular (T_3) $\implies$ Hausdorff (T_2) $\implies$ T_1



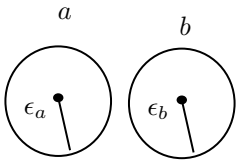
Example 6.126. *The space $\mathbb{R}_\ell$ is normal but not metrizable. On the other hand, $\mathbb{R}_\ell^2$ is regular but not normal. (For further details, see Munkres, section 31, which contains these examples and more.)*

Theorem 6.127.
- *Regular + second-countable $\implies$ normal*
- *Hausdorff + compact $\implies$ normal*

These results will not be proven here, but you can refer to Munkres, section 32, for the proofs. However, when we proved that compact subsets of Hausdorff spaces are closed, we established that compactness and Hausdorffness together imply regularity. The result on normality was covered in Homework 2.

Theorem 6.128. *Every metric space is normal.*

Proof. Let A and B be disjoint closed sets in a metric space (X, d). For each $a \in A$, there exists $\epsilon_a > 0$ such that the open ball $B_{\epsilon_a}(a) \subset X \setminus B$. Similarly, for each $b \in B$, there exists $\epsilon_b > 0$ such that $B_{\epsilon_b}(b) \subset X \setminus A$.



Now, define the sets:

$$U = \bigcup_{a \in A} B_{\epsilon_a/2}(a) \quad \text{and} \quad V = \bigcup_{b \in B} B_{\epsilon_b/2}(b)$$

Clearly, both U and V are open sets, and we have $A \subset U$ and $B \subset V$.

We claim that $U \cap V = \emptyset$. Suppose, for the sake of contradiction, that $z \in U \cap V$. Then there exist points $a \in A$ and $b \in B$ such that $z \in B_{\epsilon_a/2}(a)$ and $z \in B_{\epsilon_b/2}(b)$. By the triangle inequality, we have:

$$d(a, b) \leq d(a, z) + d(z, b) \leq \frac{\epsilon_a}{2} + \frac{\epsilon_b}{2} \leq \max(\epsilon_a, \epsilon_b).$$

This contradicts the assumption that a and b are disjoint, since if $d(a, b) < \epsilon_a$, the ball $B_{\epsilon_a}(a)$ would not be contained in $X \setminus B$ (and similarly for $B_{\epsilon_b}(b)$ not being contained in $X \setminus A$). Hence, $U \cap V = \emptyset$, as required.

□

We can now explore which topological spaces are metrizable. We have already established that metrizable spaces are first-countable and normal. However, the converse does not hold, as the space $\mathbb{R}_\ell$ provides a counterexample.

Theorem 6.129 (Urysohn's Metrization Theorem)**.** *If X is regular and has a countable basis, then X is metrizable.*

(Note: The first condition is necessary, while the second condition is stronger than required. A sharper criterion is given by the Nagata-Smirnov theorem, but it is more technical to state and prove.)

6.17 Urysohn's Lemma

Urysohn's Lemma plays a crucial role in the proof of the metrization theorem.

Theorem 6.130. *Let X be a normal space, and let A and B be disjoint closed subsets of X. Then, there exist continuous functions $f : X \to [0, 1]$ such that $f(x) = 0$ for all $x \in A$ and $f(x) = 1$ for all $x \in B$.*

Idea:

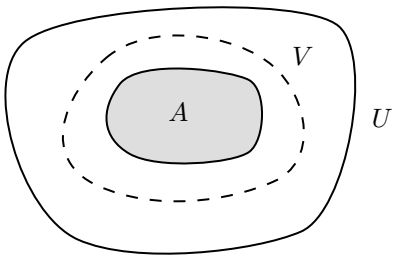
1. Construct open sets U_q for each $q \in [0, 1] \cap \mathbb{Q}$ such that $A \subset U_0 \subset \cdots \subset U_1 = X \setminus B$, with the additional property that for $p < q$, we have $\overline{U}_p \subset U_q$. Also, set $U_q = X$ for $q > 1$.

2. Define $f(x) = \inf\{q \in \mathbb{Q} \mid x \in U_q\}$, and show that f is continuous.

Step 1 relies on the following reformulation of normality.

Lemma 6.131. *Let X be normal. Then for any closed set A and any open set U containing A, there exists an open set V such that $A \subset V$ and $\overline{V} \subset U$.*

Proof. Let A and $B = X \setminus U$ be disjoint closed sets. Since X is normal, there exist open sets $V \supset A$ and $V' \supset B$ such that $V \cap V' = \emptyset$. Moreover, since $X \setminus V'$ is closed, we have $V \subset X \setminus V'$, so $\overline{V} \subset X \setminus V'$. Therefore, $A \subset V \subset \overline{V} \subset X \setminus V' \subset X \setminus B = U$.

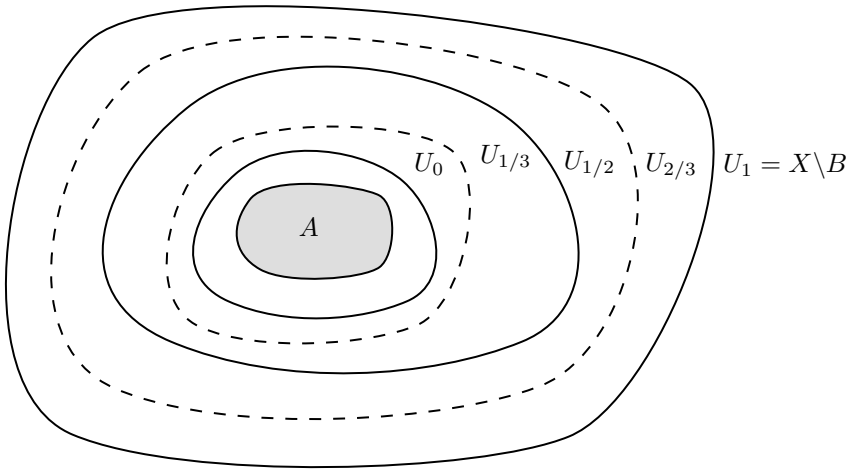


□

Proof. **Step 1**. Let A and B be disjoint closed sets. Define $U_1 = X \setminus B$ and choose an open set U_0 such that $A \subset U_0 \subset \overline{U}_0 \subset U_1$. We now construct the open sets U_q, where $q \in (0,1) \cap \mathbb{Q}$, by induction, ensuring that $p < q \implies \overline{U}_p \subset U_q$.

Label $(0,1) \cap \mathbb{Q}$ as $\{q_0, q_1, q_2, \dots\}$, with $q_0 = 0$ and $q_1 = 1$, and proceed with the induction. Suppose that $U_{q_0}, \dots, U_{q_n}$ have already been constructed. We now construct $U_{q_{n+1}}$ using the lemma above. Let $q_k = \max(\{q_0, \dots, q_n\} \cap (0, q_{n+1}])$ and $q_\ell = \min(\{q_0, \dots, q_n\} \cap (q_{n+1}, 1])$, so that $q_k < q_{n+1} < q_\ell$, with no rationals between them.

By the induction hypothesis, $\overline{U}_{q_k} \subset U_{q_\ell}$, and using normality, there exists an open set V such that $\overline{U}_{q_k} \subset V \subset \overline{V} \subset U_{q_\ell}$. Thus, we set $U_{q_{n+1}} = V$. By induction, we can construct all the U_q's, and the property $p < q \implies \overline{U}_p \subset U_q$ is satisfied. Set $U_q = \emptyset$ for $q < 0$ and $U_q = X$ for $q > 1$, which still satisfies the condition $p < q \implies \overline{U}_p \subset U_q$.

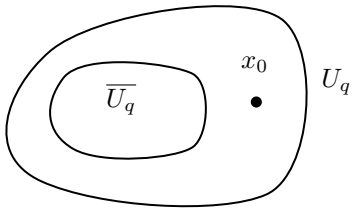


Step 2. Define $f(x) = \inf Q_x$, where $Q_x = \{q \in \mathbb{Q} \mid x \in U_q\}$. Since $U_{<0} = \emptyset$ and $U_{>1} = X$, it follows that $(1, \infty) \subset Q_x \subset [0, \infty)$ for all $x \in X$, so $f(x) \in [0, 1]$ for all $x \in X$. Moreover, if $x \in A \subset U_0$, then $f(x) = 0$, and if $x \in B$, then $x \in U_1 = X \setminus B$, so $Q_x = (1, \infty)$ and $f(x) = 1$. It remains to show that $f : X \to [0, 1]$ is continuous.

To prove continuity, observe the following:

- If $x \in \overline{U}_q$, then $f(x) \leq q$: if $x \in \overline{U}_q$, then $x \in U_{q'}$ for all $q' > q$, so $Q_x \supset \{q \mid q > q'\}$.
- If $x \notin U_q$, then $f(x) \geq q$: if $x \notin U_q$, then $Q_x \subset \{q \mid q \in (q, \infty)\}$.

Now, for any open interval (c, d), we show that $f^{-1}((c, d))$ is open in X. Assume $x_0 \in f^{-1}((c, d))$, and let $p, q \in \mathbb{Q}$ such that $c < p < f(x_0) < q < d$. By the above observations, $x_0 \in U_q$ and $x_0 \notin \overline{U}_p$. Define $V = U_q \cap (X \setminus \overline{U}_p)$, which is open and contains x_0.



Moreover, for $x \in V$:

- $x \notin U_p$ implies $f(x) \geq p$, so $V \subset f^{-1}([p, q]) \subset f^{-1}((c, d))$.

- $x \in \overline{U}_q$ implies $f(x) \leq q$, so $f^{-1}((c,d)) \supset V$, establishing that $f^{-1}((c,d))$ is open.

□

Next, we prove the metrization theorem, which states that if X is a normal space with a countable basis, then X is metrizable. We prove this by embedding X as a subspace of a metric space, namely $[0,1]^\omega$ with the product topology (or uniform topology, which is metrizable).

On the product topology, define the metric d on $[0,1]^\omega$ by

$$d((x_n),(y_n)) = \sup_n \left(\frac{1}{n}|x_n - y_n|\right),$$

and the corresponding basis of open sets is $B_\epsilon((x_n)) = \prod_n (x_n - n\epsilon, x_n + n\epsilon)$. Notably, for $n > \epsilon^{-1}$, this basis covers all of $[0,1]$.

Lemma 6.132 (Step 1)**.** *There exists a countable collection of continuous functions $f_n : X \to [0,1]$ such that for every $x_0 \in X$ and every neighborhood $U \ni x_0$, there exists some n such that $f_n(x_0) > 0$ and $f_n \equiv 0$ on $X \setminus U$.*

Proof. This result follows from Urysohn's lemma, but we need to be careful to ensure that countably many functions suffice. Let $\mathcal{B} = \{B_n\}$ be a countable basis for the topology of X. If $x_0 \in U$, where U is open, then there exists some $B_n \in \mathcal{B}$ such that $x_0 \in B_n \subset U$.

Since X is normal, there exists an open set V such that $x_0 \in V \subset \overline{V} \subset B_n$. Additionally, there exists $B_m \in \mathcal{B}$ such that $x_0 \in B_m \subset V$, which implies that $x_0 \in \overline{B_m} \subset B_n \subset U$.

Now, for every pair $(m,n) \in \mathbb{Z}_+ \times \mathbb{Z}_+$ such that $\overline{B_m} \subset B_n$, we apply Urysohn's lemma to obtain a continuous function $g_{m,n} : X \to [0,1]$ such that:

$$g_{m,n} = 1 \text{ on } \overline{B_m}, \quad g_{m,n} = 0 \text{ on } X \setminus B_n.$$

The countable collection of functions $\{g_{m,n}\}_{(m,n)\in\mathbb{Z}_+\times\mathbb{Z}_+}$ has the desired properties.

□

Lemma 6.133 (Step 2)**.** *Define $F : X \to [0,1]^\omega$ by $F(x) = (f_1(x), f_2(x), \dots)$. Then F is an embedding, i.e., it is continuous, injective, and X is homeomorphic to $F(X) \subset [0,1]^\omega$. This will show that the topology on X is the same as the subspace topology induced from the metric on $[0,1]^\omega$.*

Proof. **Continuity**: The map F is continuous in the product topology because each component f_n is continuous from $X \to [0,1]$. Since the product topology on $[0,1]^\omega$ is the coarsest topology making all the coordinate projections continuous, and each f_n is continuous, it follows that F is continuous.

Injectivity: We now show that F is injective. Suppose $x \neq y \in X$. Since x and y are distinct, there exist disjoint open sets $U \ni x$ and $V \ni y$ in X. By the properties of the functions f_n, we can find m, n such that:

- $f_n(x) > 0$, and $f_n = 0$ outside of U (hence at y),
- $f_m(y) > 0$, and $f_m = 0$ outside of V (hence at x).

Thus, $F(x) \neq F(y)$, proving that F is injective.

Homeomorphism: We now show that F is a homeomorphism from X onto $F(X) \subset [0,1]^\omega$. Since F is continuous and injective, it is a bijection between X and $F(X)$. It remains to prove that F is an open map, i.e., for every open set $U \subset X$, the image $F(U) \subset F(X)$ is open.

Let $U \subset X$ be any open set and let $x_0 \in U$. Then, there exists n such that $f_n(x_0) > 0$ and $f_n = 0$ outside of U. Define the open set in $F(X)$ as:

$$V_n = \pi_n^{-1}((0,\infty)) \cap F(X) = \{z = (z_1, z_2, \dots) \in F(X) \mid z_n > 0\}.$$

Since $f_n(x_0) > 0$, we have $x_0 \in F^{-1}(V_n) \subset U$, implying that $F(x_0) \in V_n \subset F(U)$.

This argument holds for all $x_0 \in U$, which implies $F(U)$ is open in $F(X)$. Therefore, F is an open map, and since it is a continuous bijection, it follows that F is a homeomorphism.

Thus, X is homeomorphic to $F(X) \subset [0,1]^\omega$, and since $[0,1]^\omega$ is metrizable, we conclude that X is metrizable.

□

6.18 Gluing and Quotients

One effective way to construct interesting topological spaces is by "gluing" together simpler spaces.

Example 6.134.

$$[0,1] \to S^1 \to [0,1] \times [0,1] \to [0,1] \times S^1 \to S^1 \times S^1.$$

The construction underlying this process is the **quotient topology**.

Definition 6.135. *Let X be a topological space, A a set, and $f : X \to A$ a surjective map. The **quotient topology** on A is defined as follows: a subset $U \subset A$ is open if and only if $f^{-1}(U) \subset X$ is open.*

Exercise 6.136. *Verify that this indeed defines a topology on A, and that it is the finest topology on A such that f is continuous.*

Definition 6.137. *A map $f : X \to Y$ between topological spaces is called a **quotient map** if f is surjective, and a subset $U \subset Y$ is open if and only if*

$f^{-1}(U) \subset X$ is open. In other words, the topology on Y is the quotient topology induced by the map $f : X \to Y$.

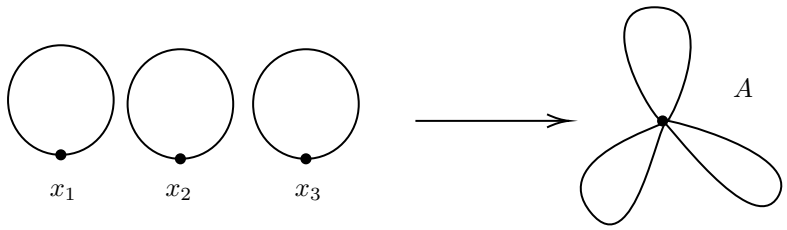
Typically, we start with an equivalence relation $\sim$ on X, define $A = X/\sim$ to be the set of equivalence classes, and define the map $f : X \to X/\sim = A$ by $f(x) = [x]$. Conversely, given any surjective map $f : X \to A$, we can define an equivalence relation on X by $x \sim x' \iff f(x) = f(x')$, so that $X/\sim = A$.

Example 6.138. *Consider $S^1 \simeq [0,1]$ with the endpoints 0 and 1 glued together. This is done by setting $0 \sim 1$, so that the set $\{0,1\}$ forms a single equivalence class (while every other point remains a singleton equivalence class). The quotient map is given by $f : [0,1] \to S^1$, where*

$$f(t) = (\cos(2\pi t), \sin(2\pi t)).$$

One should check that away from the endpoints, f is a homeomorphism. That is, for $t \in (0,1)$, $f(t)$ maps to $S^1 \setminus \{(0,1)\}$, so the only points that require checking are at $t = 0$ and $t = 1$. Specifically, for any open set $U \ni (1,0)$ in S^1, we have $f^{-1}(U) \supset \{0,1\}$, which is open in $[0,1]$. This contrasts with the map $g = f|_{[0,1)} : [0,1) \to S^1$, which is not a quotient map. For instance, if $V = g([0,\epsilon))$ is open in S^1, then $g^{-1}(V) = [0,\epsilon)$ is open in $[0,1)$, but V is not open in S^1, which would contradict the requirements for g to be a quotient map. However, for f, $f^{-1}(V) = [0,\epsilon) \cup \{1\}$, which is open in $[0,1]$.

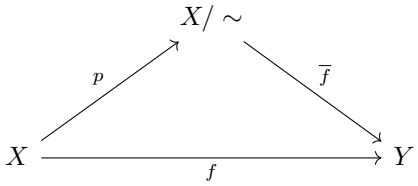
Example 6.139. *Let $X_1, \ldots, X_n$ be topological spaces, each homeomorphic to S^1, and let $x_i \in X_i$ be a basepoint for each i. Let A be the quotient space of the disjoint union $\bigsqcup_{i=1}^n X_i$ by the equivalence relation $x_i \sim x_j$ for all i, j. This space is called the **wedge** of n circles, often denoted $\bigvee_{i=1}^n S^1$, which is the result of gluing the circles at their respective basepoints:*



There is a useful characterization of continuous maps from a quotient space. Suppose $A = X/\sim$ and $f : X \to Y$ is a map such that $x \sim x' \implies f(x) = f(x')$. In this case, we can define a map $\overline{f} : X/\sim \to Y$ by setting

$$\overline{f}([x]) = f(x),$$

where $[x]$ denotes the equivalence class of x in $X/\sim$:



Theorem 6.140. *If $f : X \to Y$ is a continuous map and $x \sim x' \implies f(x) = f(x')$, then equipping $X/\sim$ with the quotient topology, the map $\overline{f} : X/\sim \to Y$ is continuous.*

Proof. Let $p : X \to X/\sim$, $p(x) = [x]$, be the quotient map. Recall that $\overline{f}([x]) = f(x)$ for any $x \in [x]$, and hence $\overline{f} \circ p = f$.

Now, let $U \subset Y$ be an open set. Since f is continuous, $f^{-1}(U)$ is open in X. We have

$$f^{-1}(U) = p^{-1}(\overline{f}^{-1}(U)),$$

which is open in X. By the definition of the quotient topology, we know that $V \subset X/\sim$ is open if and only if $p^{-1}(V)$ is open in X. Therefore, we conclude that $\overline{f}^{-1}(U) \subset X/\sim$ is open.

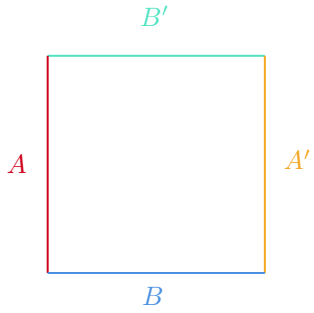
□

Thus, since $p : X \to X/\sim$ is continuous: $\overline{f}$ is continuous if and only if $f = \overline{f} \circ p$ is continuous.

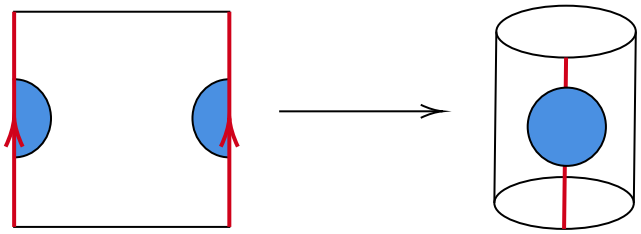
Example 6.141. *Let $X = \mathbb{R}^{n+1} \setminus \{0\}$. Define an equivalence relation $x \sim y$ if and only if x and y lie on the same line through the origin, i.e., $x = \alpha y$ for some $\alpha \in \mathbb{R} \setminus \{0\}$. This defines an equivalence relation, and the quotient space is the projective n-space $\mathbb{RP}^n = X/\sim$, equipped with the quotient topology. This space can be interpreted as the space of lines through the origin in $\mathbb{R}^{n+1}$.*

If Y is another topological space, then a continuous map $\overline{f} : \mathbb{RP}^n \to Y$ is equivalent to a continuous map $f : \mathbb{R}^{n+1} \setminus \{0\} \to Y$ that satisfies $f(\alpha x) = f(x)$ for all $\alpha \in \mathbb{R} \setminus \{0\}$ and $x \in X$. (More details about $\mathbb{RP}^n$ can be found in the homework.)

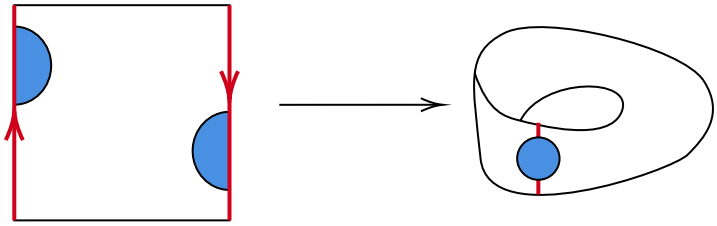
Example 6.142. *Consider various quotients of the unit square $X = [0,1]^2$. Let the edge $A = \{0\} \times [0,1]$, and define equivalence relations involving the edges A', B, and B', as follows:*



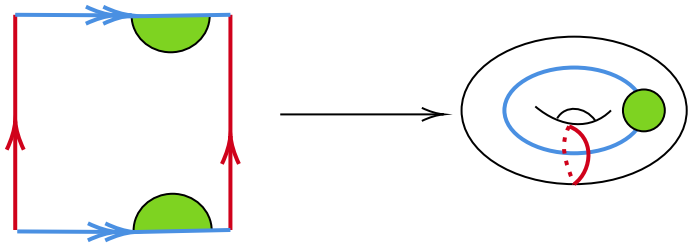
1. *Gluing A to A' by $(0,t) \sim (1,t)$ results in a cylinder. A neighborhood of a point on the gluing line corresponds to two neighborhoods: one in A near $(0,t)$ and one in A' near $(1,t)$ in X.*



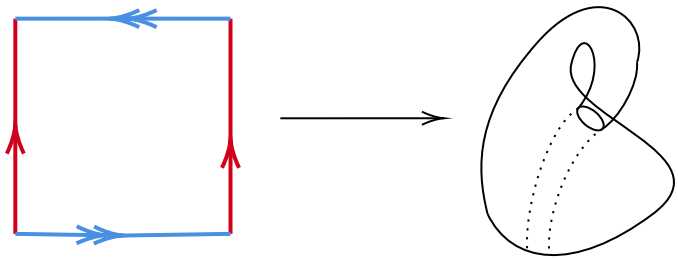
2. *If instead, we glue A to A' by $(0,t) \sim (1,1-t)$, we obtain a* ***Möbius band!***



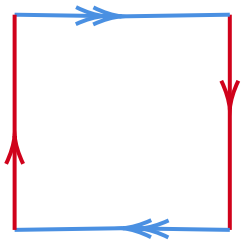
3. *Gluing A to A' via $(0,t) \sim (1,t)$ and B to B' by $(s,0) \sim (s,1)$ gives a torus.*



4. *Gluing $(0,t) \sim (1,t)$ and $(s,0) \sim (1-s,1)$, however, gives the* ***Klein bottle****, which cannot be embedded in $\mathbb{R}^3$ without self-intersection. We can draw a picture of it, but it will necessarily be a self-intersecting diagram.*



5. *Gluing* $(0,t) \sim (1,1-t)$ *and* $(s,0) \sim (1-s,1)$ *is tricky to visualize, but the resulting quotient space is actually homeomorphic to* $\mathbb{RP}^2$.



Exercise 6.143. *What happens if we glue* $(0,t) \sim (t,0)$ *and* $(1,s) \sim (s,1)$*? What shape does that form?*

7 Algebraic Topology

7.1 Homotopy

Homotopy is the notion of continuous deformation, parametrized by $I = [0, 1]$.

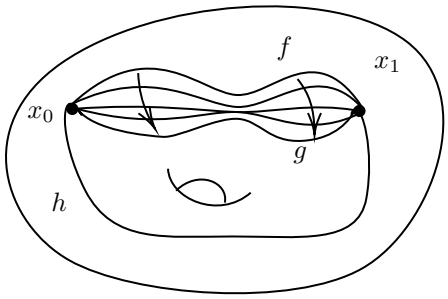
Definition 7.1. *$f, g : X \to Y$ two continuous maps. A **homotopy** between f and g is a continuous map $H : X \times I \to Y$ such that $H(x, 0) = f(x), H(x, 1) = g(x) \forall x \in X$. If this exists then say f and g are **homotopic** and write $f \sim g$. If f is homotopic to a constant map, we say it is **nullhomotopic**.*

We want to study **paths** in topological spaces, ie. $f : [0, 1] \to X$ continuous, $f(0) = x_0, f(1) = x_1$.

The above notion is not useful for paths if we don't fix the end points x_0 and x_1 (see HW4).

Better notion: **homotopy of paths** only considers homotopies which keep the end points in place.

General notion: pairs $(X, A), A \subset X$ subspace, maps of pairs $(X, A) \xrightarrow{f} (Y, B) : f(A) \subset B$.



$f \simeq_p g$ homotopic paths, h not homotopic to f and g

Definition 7.2. *Two paths $f, g : I \to X$ from x_0 to x_1 are **(path) homotopic** if $\exists$ continuous $H : I \times I \to X$ such that $H(s, 0) = f(s), H(s, 1) = g(s)$ (homotopy) and $H(0, t) = x_0, H(1, t) = x_1$ (fix end points: so $\forall t \in [0, 1], f_t = H|_{I \times t}$ is a path from x_0 to x_1). Such H is a **path homotopy**, and we write $f \sim_p g$.*

Lemma 7.3. *$\simeq_w$ and $\simeq_p$ are equivalence relations.*

Proof.

- Clearly $f \simeq f$ (constant homotopy $H(x, t) = f(x)$)
- If $f \simeq g$ with homotopy $F(x, t)$, then the **reverse homotopy** $G(x, t) = F(x, 1 - t)$ gives $g \simeq f$.

- Assume $f \simeq g$ with homotopy $F(x,t)$, $g \simeq h$ with homotopy $G(x,t)$, then the **concatenation** of these $H : X \times [0,1] \to Y$ defined by

$$H(x,t) = \begin{cases} F(x,2t) \text{ if } t \in [0,\frac{1}{2}] \\ G(x,2t-1) \text{ if } t \in [\frac{1}{2},1] \end{cases} .$$

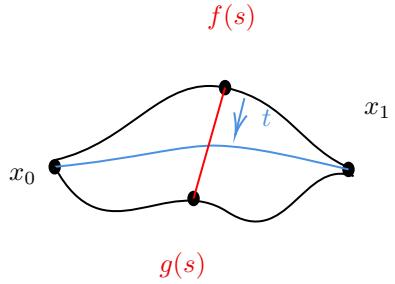
These two formulas agree at $t = \frac{1}{2}$ ($F(x,1) = g(x) = G(x,0)$) so H is well-defined and continuous ("pasting lemma" Thm 18.3) and gives a homotopy $f \simeq h$.

- In the case of path homotopies, can check the above constructions preserve the requirements $F(0,t) = x_0$ and $F(1,t) = x_1$, so yield path homotopies.

□

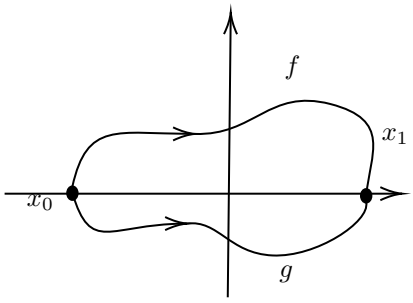
Example 7.4.

1. *If f, g are paths in $\mathbb{R}^n$ (or any convex subset of $\mathbb{R}^n$) from x_0 to x_1, we can define the **straight-line homotopy** $F(s,t) = (1-t)f(s) + tg(s)$.*



For each s, this connects $f(s)$ to $g(s)$ by a straight line segment. We conclude: $f \simeq_p g$ always!

2. *The punctured plane $X = \mathbb{R}^2 - \{(0,0)\}$, let f, g be paths from $(-1,0)$ to $(1,0)$ such that f stays in the upper half plane $\{(x,y)|y \geq 0\}$, g stays in the lower half plane $\{(x,y)|y \leq 0\}$.*



Then there is no homotopy between f and G in X (We'll prove this rigorously later).

Definition 7.5. *Spaces X, Y are **homotopy equivalent** if $\exists f : X \to Y, g : Y \to X$ such that $f \circ g \simeq id_U, g \circ f \simeq id_X$ homotopic (vs. exact inverse would be homeomorphic).*

Check: this is an equivalence relation.

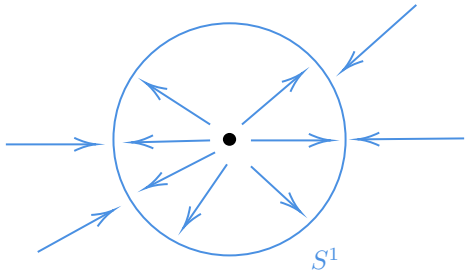
Definition 7.6. *X is contractible if X is homotopy equivalent to $\{point\}$.*

Example 7.7. *$\mathbb{R}^n$ (or a convex subset of $\mathbb{R}^n$) is contractible: ie. $\{0\} \overset{i}{\hookrightarrow} \mathbb{R}^n, r : x \mapsto 0$.*

Check: $i \cdot r =$ zero map is homotopic to $\mathrm{id}_{\mathbb{R}^n}$ by $H(x,t) = tx$.

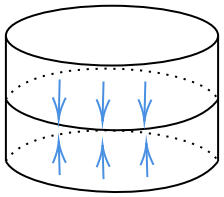
Example 7.8. *$\mathbb{R}^2 - \{0\}$ is not contractible, but homotopy equivalent to S^1, via r is a **deformation retraction** of $X = \mathbb{R}^2 - \{0\}$ onto its subset $A = S^1 \subset X$, ie.*

- $r : X \to A$
- $r|_A = id_A$ *(ie. $r \circ i = id_A$)*
- $i \circ r : X \to A \subset X$ *is $\simeq id_X$.*

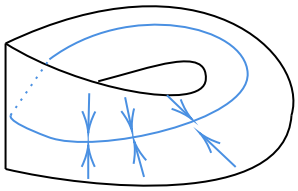


In this case, $i \circ r(x) = \frac{x}{|x|}$ homotopic to id by straight line homotopy. Deformation retraction is a useful special case of homotopy equivalence.

By the same argument, the cylinder $S^1 \times I$

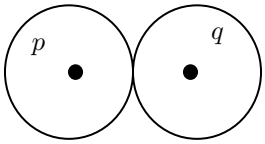


and Möbius band



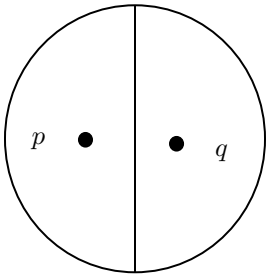
deformation retract onto "middle" S^1 by sliding points of $[0,1]$ to midpoint. (Check: this is consistent with the twisted gluing of $I \times I, (0,y) \sim (1,1-y)$.) Hence they are homotopy equivalent to S^1 (and to each other and to $\mathbb{R}^2 - \{0\}$).

Example 7.9. $\mathbb{R}^2 - \{(p,q)\}$ *deformation retracts onto wedge of two* S^1*'s ("figure 8" space).*



$S^1 \vee S^1$

Or also on "theta" graph" (homotopy equivalent to ∞*, not homeomorphic!)*

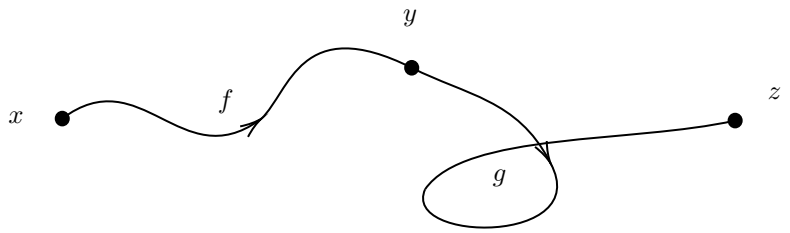


7.2 The Fundamental Group

We now focus on paths and path homotopy as a way to define an algebraic invariant of topological spaces (up to homotopy equivalence): the **fundamental group**. A group needs a multiplication?

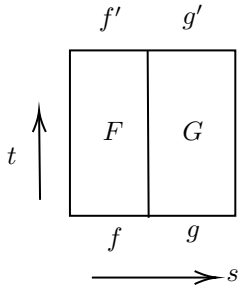
Definition 7.10. *If* f *is a path from* x *to* y *and* g *is a path from* y *to* z*, define a path* $f * g$ *from* x *to* z *running through first* f *then* g *(twice as fast):*

$$(f * g)(s) = \begin{cases} f(2s) \text{ if } s \in [0, \frac{1}{2}] \\ g(2s-1) \text{ if } s \in [\frac{1}{2}, 1] \end{cases}$$



This product is well-defined on path-homotopy classes, as long as $f(1) = g(0)$: if $f \simeq_p f'$ and $g \simeq_p g'$ then $f * g \simeq_p f' * g'$ using homotopy

$$(F * G)(s,t) = \begin{cases} F(2s,t) \text{ if } s \leq \frac{1}{2} \\ G(2s-1,t) \text{ if } s \geq \frac{1}{2} \end{cases}$$



So we define $[f] * [g] = [f * g]$.

Proposition 7.11. *This operation is associative, and has identity and inverses.*

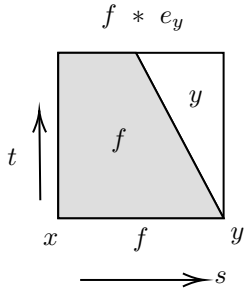
There is also the "**fundamental groupoid**" of X: the category with objects = points of X and $\mathrm{Mor}(x,y) = \{\text{path homotopy classes of paths } s \to y\}$.

Remark 7.12. *Category: composition is associative and $\exists$ identity morphisms $x \to x$.*

Groupoid: all morphisms have inverses.

Now let's prove the proposition.

Proof. **Identity**: given $x \in X$, consider the constant path $e_x : I \to X, e_x(s) = x \forall s$, and let $\mathrm{id}_x = [e_x]$. We claim that if f is any path from x to y, then $[f] * \mathrm{id}_y = \mathrm{id}_x * [f] = [f]$. Indeed, there are explicit homotopies $f \simeq_p (f * e_y)$ and similarly, $(e_x * f) \simeq_p f$.



This is through

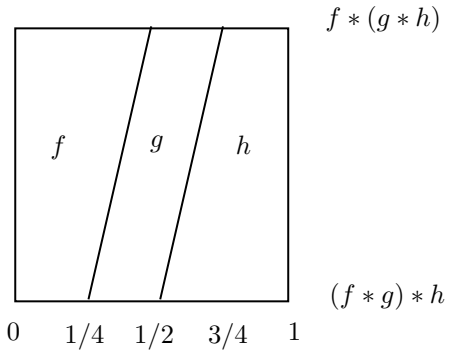
$$F(s,t) = \begin{cases} f\left(\frac{s}{1-\frac{t}{2}}\right) \text{ if } s \in [0, 1-\frac{t}{2}] \\ y \text{ if } s \in [1-\frac{t}{2}, 1] \end{cases}$$

Inverse: given a path f from x to y, define the reverse path $\overline{f}(s) = f(1-s)$ from y to x. $[\overline{f}]$ is inverse to $[f]$, namely $e_x \simeq_p f * \overline{f}$ and $e_y \simeq_p \overline{f} * f$. Indeed:

$$F(s,t) = \begin{cases} f\left(2ts\right) \text{ if } s \in [0, \frac{t}{2}] \\ f(2t(1-s)) \text{ if } s \in [\frac{t}{2}, 1] \end{cases}$$

For given t, this runs forward along f from $f(x) = 0$ to $f(t)$ at $s = \frac{1}{2}$, then backwards to $f(0) = x$ at $s = 1$. For $t = 0$ get e_x, for $t = 1$ get $f * \overline{f}$. (Similarly for $e_y \simeq_p \overline{f} * f$).

Associativity: Given paths f, g, h with $f(1) = g(0)$ and $g(1) = h(0)$, claim $(f * g) * h \simeq_p f * (g * h)$. Both run along f then g then h, but with different parameterizations. The homotopy comes from adjusting for this:



Let

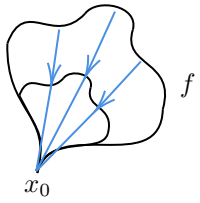
$$F(s,t) = \begin{cases} f\left(\frac{4s}{1+t}\right) & \text{if } s \in [0, \frac{1+t}{4}] \\ g(4s-(1+t)) & \text{if } s \in [\frac{1+t}{4}, \frac{2+t}{4}] \\ h\left(\frac{4s-(2+t)}{2-t}\right) & \text{if } s \in [\frac{2+t}{4}, 1] \end{cases}$$

□

Now let's finally talk about the fundamental group. Groups are much easier to study than groupoids! We want to be able to multiply always, not worrying whether end points match. THus we fix a **base point** $x_0 \in X$ and only consider paths from x_0 to itself - ie. **loops** (based at x_0).

Definition 7.13. *The set of path homotopy classes of loops based at x_0, with operation $*$ (concatenation), is called the **fundamental group** of X, denoted $\pi_i(X, x_0)$.*

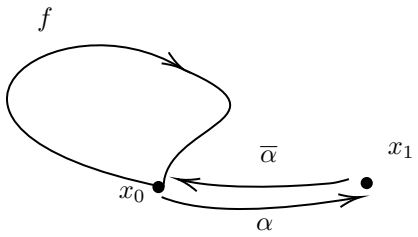
Example 7.14. *In $\mathbb{R}^n$ (or a convex domain in $\mathbb{R}^n$), every loop at x_0 is path homotopic to the identity (ie. the constant path at x_0) by the straight-line homotopy $F(t,s) = (1-t)f(s) + tx_0$. So $\pi_1(\mathbb{R}^n, x_0) = \{id\}$.*



Definition 7.15. *X is **simply-connected** if $X \neq$ is path-connected, and for $x_0 \in X, \pi_1(X, x_0) = \{1\}$.*

This definition is sensible because π_1 is, up to isomorphism, independent of choice of x_0 inside a path component of X (we'll see this next time).

Let's discuss the dependence on the base point. If x_0, x_1 are in the same path-component of X, let α be a path from x_0 to x_1. Then for any loop f based at x_0, we get a loop at x_1 by taking $\overline{\alpha} * f * \alpha$ and so we get a map $\hat{\alpha} : \pi_1(X, x_0) \to \pi_1(X, x_1), [f] \mapsto [\overline{\alpha} * f * \alpha] = [\overline{\alpha}] * [f] * [\alpha]$. (Recall that $*$ is well-defined on path-homotopy classes).



Proposition 7.16. *$\hat{\alpha} : \pi_1(X, x_0) \to \pi_1(X, x_1)$ is a **group isomorphism**.*

Proof.

- If $a, b \in \pi_1(X, x_0)$ then

$$\begin{aligned}\hat{\alpha}(a * b) &= [\hat{\alpha}]^{-1} * (a * b) * [\alpha] \\ &= [\overline{\alpha}] * a * [\alpha] * [\overline{\alpha}] * b * [\alpha] \\ &= \hat{\alpha}(a) * \hat{\alpha}(b) \\ &= \hat{\alpha}(a) * \hat{\alpha}(b)\end{aligned}$$

  so $\hat{\alpha}$ is a group homomorphism.

- Let $\beta = \hat{\alpha}$ reverse path from x_1 to x_0. Then $\hat{\beta} : \pi_1(X, x_1) \to \pi_1(X, x_0)$. We claim $\hat{\beta}$ and $\hat{\alpha}$ are inverses of each other. Indeed: for $a \in \pi_1(X, x_0)$,

$$\begin{aligned}\hat{\beta}(\hat{\alpha}(a)) &= \hat{\beta}([\overline{\alpha}] * a * [\alpha]) \\ &= [\overline{\beta}] * [\overline{\alpha}] * a * [\alpha] * [\beta] \\ &= [\alpha] * [\overline{\alpha}] * a * [\alpha] * [\overline{\alpha}] \\ &= a\end{aligned}$$

  Hence $\hat{\beta} \circ \hat{\alpha} = \mathrm{id}$ (and similarly $\hat{\alpha} \circ \hat{\beta} = \mathrm{id}$ as well), so $\hat{\alpha}$ is an isomorphism.

□

Corollary 7.17. *If X is path-connected, then $\pi_1(X, x_0)$ is independent of x_0 up to isomorphism.*

Remark 7.18. *When α is a loop at x_0, we get an **automorphism** $\hat{\alpha}$ of $\pi_1(X, x_0)$. This is in fact an inner automorphism = conjugation by $[\alpha]$: $a \mapsto [\alpha]^{-1} * a * [\alpha]$.*

Let's consider π_1 as a functor: Consider the category of **pointed topological spaces**:

- Objects = topological space + choice of base point, (X, x_0)
- Morphisms = continuous maps preserving base points: $f : (X, x_1) \to (Y, y_0)$ means $f : X \to Y$ continuous and such that $f(x_0) = y_0$.

Proposition 7.19. *A continuous map $h : (X, x_0) \to (Y, y_0)$ induces a **group homomorphism** $h_* : \pi_1(X, x_0) \to \pi_1(Y, y_0)$ defined by $h_*([f]) = [h \circ f]$.*

Check:

- If $f \simeq_p f'$ via F then $h \circ f \simeq_p h \circ f'$ via $h \circ F$. So h_* is well-defined.
- $h \circ (f * g) = (h \circ f) * (h \circ g)$ (composition with h compatible with concatenation). So h_* is a group homomorphism, $h_*([f] * [g]) = h_*([f]) * h_*([g])$.

Proposition 7.20. *Given* $(X, x_0) \xrightarrow{h} (Y, y_0) \xrightarrow{k} (Z, z_0)$, $(k \circ h)_* = k_* \circ h_* : \pi_1(X, x_0) \to \pi_1(Z, z_0)$. *Hence* π_1 *is a* ***functor*** *(maps composition* $k \circ h$ *to composition* $k_* \circ h_*$*).*

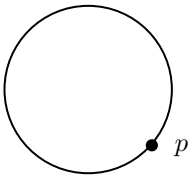
(This is just $(k \circ h) \circ f = k \circ (h \circ f)$). This implies the following:

Corollary 7.21. *If* $h : (X, x_0) \to (Y, y_0)$ *is a homeomorphism, then* h_* *is an isomorphism.*

But we can do better! Recall:

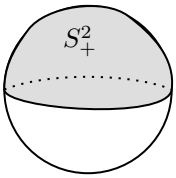
- A **retraction** of X onto a subset $A \hookrightarrow X$ is $r : X \to A$ such that $r|_A = \mathrm{id}_A$, ie. $r \circ i = \mathrm{id}_A$. Then, taking a base point $a_0 \in A$, $\pi_1(A, a_0) \underset{\leftarrow r_*}{\overset{i_*}{\rightleftarrows}} \pi_1(X, a_0)$ and $r_* \circ i_* = \mathrm{id} \implies \mathrm{Ker}(i_*) = \{1\}$, ie. i_* injective.
- A **deformation retraction** = assume moreover that $i \circ r : X \to X$ is homotopic to id_X by a homotopy that fixes A. Then we claim i_*, r_* are inverse isomorphisms, $\pi_1(A, a_0) \simeq \pi_1(X, a_1)$.

Example 7.22.



$S^1 \to p$

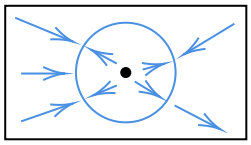
constant map



$S^2 \to S^2_+$

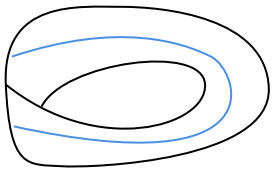
$(x, y, z) \mapsto (x, y, |z|)$

Retractions $i \circ r \not\simeq id_X$.



$\mathbb{R} - \{0\} \to S^1$

$x \mapsto x/|x|$



Mobius band $\to S^1$

are deformation retractions

More generally, recall a **homotopy equivalence** if $X \underset{\leftarrow g}{\overset{f}{\rightleftarrows}} Y$ such that $f \circ g \simeq \mathrm{id}_Y, g \circ f \simeq \mathrm{id}_X$.

Theorem 7.23. *Homotopy equivalences induce isomorphisms* $\pi_1(X, x_0) \overset{\simeq}{\underset{f_*}{\rightarrow}} \pi_1(Y, f(x_0))$.

This follows from the fact that homotopic maps induce the same homeomorphisms on π_1, namely:

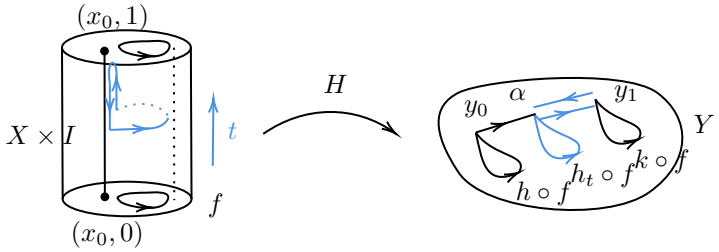
Proposition 7.24. *1. Let $h, k : X \to Y$ homotopic via a homotopy $H : X \times I \to Y$ such that $H(x_0, t) = y_0 \forall t$. Then $h_* = k_* : \pi_1(X, x_0) \to \pi_1(Y, y_0)$.*

2. If the homotopy H doesn't fix base points, let α be the path $y_0 \to y_1$ defined by $\alpha(t) = H(x_0, t) = y_t$. Then $h_* : \pi_1(X, x_0) \to \pi_1(Y, y_0), k_* : \pi_1(X, x_0) \to \pi_1(Y, y_1)$ ($k_* = \hat{\alpha} \circ h_*$)are related by the isomorphism $\hat{\alpha} : \pi_1(Y, y_0) \to \pi_1(Y, y_1)$.

Proof.

1. Given a loop $f : I \to X$ based at x_0, $I \times I \overset{f \times \mathrm{id}}{\to} X \times I \overset{H}{\to} Y, (s, t) \mapsto (f(s), t) \mapsto H(f(s), t)$. $H \circ (f \times \mathrm{id}) : I \times I \to Y$ gives a path homotopy (based at y_0) $h \circ f \simeq_p k \circ f$, hence $h_*([f]) = k_*([f])$.
2. Now consider $I \times I \vec{F} X \times I$ defined by concatenating (path $(x_0, 1) \to (x_0, t)$, loop f in $X \times \{t\}$, path $(x_0, t) \to (x_0, 1)$) then $H \circ F$ is a path homotopy in (Y, y_1) from $\alpha^{-1} * (h \circ f) * \alpha$ to $e * (k \circ f) * e$.

□



F path homotopy in$(X \times I, (x_0, 1))$

Now let's prove the theorem.

Proof. If $(X, x_0) \overset{f}{\to} (Y, y_0) \overset{g}{\to} (X, x_1)$ homotopy inverses $g \circ f \simeq \mathrm{id}_X$. By the proposition,

$$\pi_1(X, x_0) \xrightarrow{f_*} \pi_1(Y, y_0) \xrightarrow{g_*} \pi_1(X, x_1) \xrightarrow{f_*^1} \pi_1(Y, y_1)$$

with $(f \circ g)_*$ over the whole chain and $(g \circ f)_*$ under it.

where $(g \circ f)_* = \hat{\alpha}$ for some path $\alpha : x_0 \to x_1$, which is an isomorphism. Hence f_* is injective and g_* is surjective. Similarly, $(f \circ g)_*$ isomorphic to $\pi_1(Y, y_0) \to \pi_1(Y, y_1) \implies g_*$ injective, f_*^1 surjective. Hence g_* is an isomorphism, and $f_* = (g_*)^{-1} \circ \hat{\alpha}$ is also an isomorphism

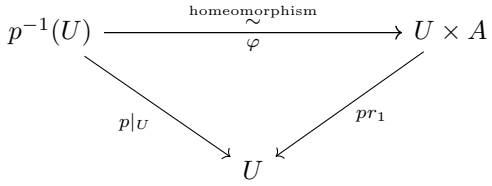
□

7.3 Covering Spaces

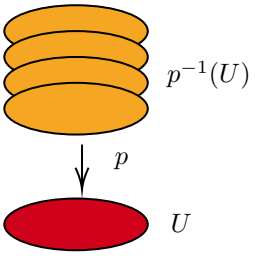
At some point we'd like to show $\pi_1(S^1) \equiv \mathbb{Z}$. We'll do this by introducing a key tool for the study of π_1: the notion of covering spaces.

Definition 7.25. *Let $p : E \to B$ be a continuous surjective map. We say p **evenly covers** an open subset $U \subset B$ if $p^{-1}(U) = \bigcup_{\alpha \in A} V_\alpha$ where $V_\alpha \subset E$ are disjoint open subsets, and for each $\alpha \in A$, $p|_{V_\alpha} : V|_\alpha \to U$ is a homeomorphism. The V_α are called **slices**.*

Equivalently, there exists

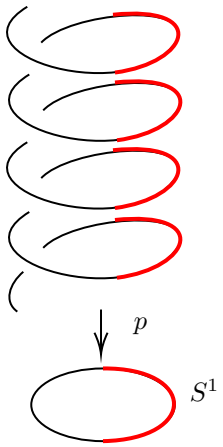


under the discrete topology such that $p|_U = pr_1 \circ \varphi$.



Definition 7.26. *If every open point of B has a neighborhood which is evenly covered by p, we say E is a **covering space** of B and p is a **covering map**. B is called the **base** of the covering.*

Example 7.27. *Define $p : \mathbb{R} \to S^1, p(t) = (\cos t, \sin t)$. This is a covering map! For instance consider $(1, 0) \in S^1$ and the neighborhood $U = \{(x, y) \in S^1 | x > 0\}$. Then $p^{-1}(U) = \bigsqcup_{n \in \mathbb{Z}} (2\pi n - \frac{\pi}{2}, 2\pi n + \frac{\pi}{2})$ and p is a homeomorphism on each slice.*

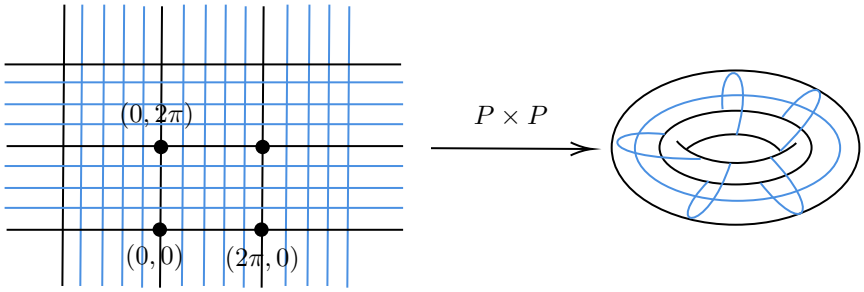


Theorem 7.28. *$p : E \to B, q : E' \to B'$ covering maps $\implies$ $p \times q : E \times E' \to B \times B'$ is a covering map.*

Proof. Given $(b, b') \in B \times B'$, let $U \ni b$ and $U' \ni b'$ be neighborhoods such that $p^{-1}(U) = \bigsqcup V_a, q^{-1}(U') = \bigsqcup V_B$ slices, then $(p \times q)^{-1}(U \times U') = p^{-1}(U) \times q^{-1}(U') = \bigsqcup_{\alpha,\beta} V_\alpha \times V'_\beta$ union of open slices homeomorphic to $U \times U'$.

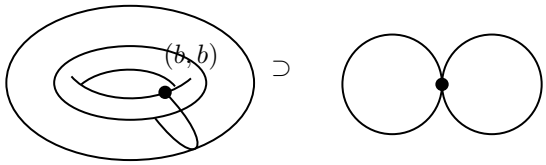
□

Example 7.29. *Consider the torus $S^1 \times S^1$ Since $\mathbb{R}$ covers S^1, $\mathbb{R}^2$ covers $S^1 \times S^1$.*

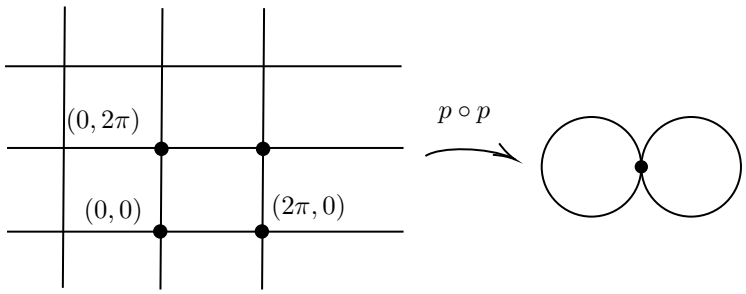


If $p : E \to B$ is a covering, and $B_0 \subset B$ is a subspace, then by restriction we get a covering $p^{-1}(B_0) \to B_0$.

Example 7.30. *For $b \in S^1$ base point on the circle, let $B_0 = (b \times S^1) \cup (S^1 \times b) \subset S^1 \times S^1$, the "figure eight space."*

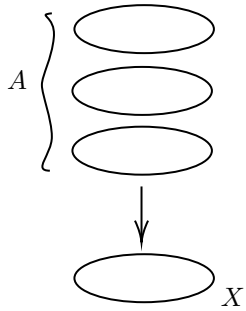


Then we have a covering $(p \circ p^{-1})^{-1}(B_0) \to B_0, (p \times p)^{-1}(B_0) = (\mathbb{R} \times 2\pi\mathbb{Z}) \cup (2\pi\mathbb{Z} \times \mathbb{R}) \subset \mathbb{R}^2$.

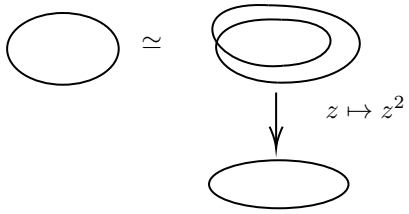


where the horizontal lines map to the left circle and the vertical lines map to the right circle.

Example 7.31. *If X any topological space, A a set with discrete topology, then $p_1 : X \times A (\simeq \sqcup_{\alpha \in A} X \times \{\alpha\}) \to X$ is a covering map.*

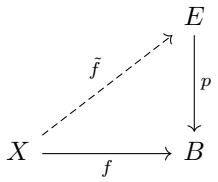


Example 7.32. *Consider $S^1 = \{z \in \mathbb{C} |\ |z| = 1\}$, then $p : S^1 \to S^1, z \mapsto z^n$ (so $e^{i\theta} \mapsto e^{in\theta}$) is an n-fold covering.*



7.4 Lifting

Definition 7.33. *Given $p : E \to B$ continuous map, a **lifting** of a continuous map $f : X \to B$ is a map $\widetilde{f} : X \to E$ such that $p \circ \widetilde{f} = f$, ie.*

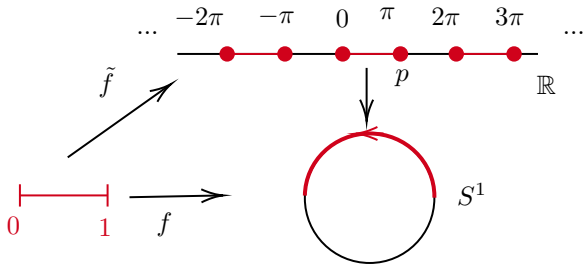


commutes.

If $p : E \to B$ is a covering map, then we can locally life, namely if $f(X) \subset U \subset B$ and U is evenly covered, then we can lift f to one of the sheets.

Key point: if $p : E \to B$ covering then paths and path homotopies in B always lift.

Example 7.34. *Consider $p : \mathbb{R}^2 \to S^1, p(x) = (\cos x, \sin x)$ and the path $f(s) = (\cos \pi s, \sin \pi s) : I \to S^1$. This has infinitely many possible lifts to paths in $\mathbb{R}$, depending on where 0 gets lifted to.*



Theorem 7.35. *$p : E \to B$ covering map, $f : [0,1] \to B$ a path starting at $f(0) = b$, and $e \in p^{-1}(b)$, Then there exists a **unique** lift $\widetilde{f}$: $[0,1] \to E$ such that $\widetilde{f}\,(0) = e$.*

Proof. Cover B by open sets U_α which are evenly covered by p. Then the preimages $f^{-1}(U_\alpha)$ are an open cover of $[0,1]$, which is compact, so $\exists$ Lebesgue number $\delta > 0$ such that $\forall x, (x, x+\delta) \subset f^{-1}(U_\alpha)$ for some α. Hence we can find a finite subdivision $0 = s_0 < s_1 < ... < s_n = 1$ such that each $f([s_i, s_{i+1}])$ lies inside one of the U_α.

Define $\widetilde{f}\,(0) = e$. Assume we have defined $\widetilde{f}\,(s)$ for $s \in [0, s_i]$. Then we define $\widetilde{f}\,(s)$ for $s \in [s_i, s_{i+1}]$ as follows. Recall $f([s_i, s_{i+1}]) \subset U$ for some U which is evenly covered by p, $p^{-1}(U) = \bigsqcup$ slices. Let V be the slice which contains $\widetilde{f}\,(s_i)$. The map $p|_V : V \to U$ is a homeomorphism, so has a continuous inverse and we can define $\widetilde{f}\,(s) = p^{-1}|_V(f(s))$ for $s \in [s_i, s_{i+1}]$, which extends $\widetilde{f}$ continuously over $[s_i, s_{i+1}]$. Repeating the process, we obtain a continuous lift $\widetilde{f}$: $[0,1] \to E$.

$\widetilde{f}$ is unique since for each s_i there was a unique slice containing $\widetilde{f}(s_i)$ and a unique way to lift $f|_{[s_i,s_{i+1}]}$ into it.

□

Theorem 7.36. *Let $F : I \times I \to B$ be continuous with $F(0,0) = b, p : E \to B$ a covering map, $e \in p^{-1}(b)$, then $\exists$ unique lift $\widetilde{F}$: $I \times I \to E$ such that $\widetilde{F}(0,0) = e$.*

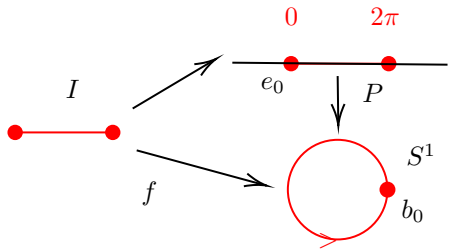
Proof. The proof is exactly the same, subdividing $I \times I$ into squares of side lengths $< \delta$ which map into open subsets of B that are evenly covered, then constructing the lift $\widetilde{F}$ one square at a time.

□

Observe: if F is a path-homotopy from f to g (in B), then $\widetilde{F}$ is a path-homotopy (in E) from $\widetilde{f}$ to $\widetilde{g}$. Indeed, if $F(0,t) = b$ for all t, then $\widetilde{F}(0,t) \in p^{-1}(b)$ which is a discrete subset of E (one point in each slice), so we must have $\widetilde{F}(0,t) = e$ for all t (always the same pint). Similarly for the other end point $\widetilde{F}(1,t)$.

On the other hand, loops don't always lift to loops! But since path-lifting is unique, given a starting point $e_0 \in p^{-1}(b_0)$, the end point is uniquely determined. This leads to a key notion:

Example 7.37. *We have:*



Definition 7.38. *The **lifting correspondence** $\varphi : \pi_1(B, b_0) \to p^{-1}(b_0)$ for a covering $(E, e_0) \xrightarrow{p} (B, b_0)$ defined by $\varphi([f]) = \widetilde{f}(1)$, where $\widetilde{f}$ is the lift of f such that $\widetilde{f}(0) = e_0$.*

Question: Why is φ well-defined? (ie. independent of choice of f in its homotopy class?)

Answer: if F is a path homotopy $f \simeq_p g$, then its lift $\widetilde{F}$ starting at e_0 is a path homotopy between $\widetilde{f}$ and $\widetilde{g}$, so $\widetilde{f}(1) = \widetilde{g}(1)$.

Example 7.39. *For the covering $p : \mathbb{R} \to S^1$. taking $b_0 = (1, 0), e_0 = 0 \in \mathbb{R}$, if f loops around the circle k times (counting counterclockwise) then its lift $\widetilde{f}$ ends at $\varphi([f]) = \widetilde{f}(1) = 2\pi k$. This gives a map $\pi_1(S^1, (1, 0)) \to 2\pi\mathbb{Z}$ (surjective.)*

Now we know, at last, that S^1 isn't simply connected?

Proposition 7.40. *If E is path connected then $\varphi : \pi_1(B, b_0) \to p^{-1}(b_0)$ is surjective.*

Proof. Let $e \in p^{-1}(b_0), g : I \to E$ a path from e_0 to e, then $f = p \circ g : I \to B$ is a loop at b_0 whose lift starting at e_0 is $\widetilde{f} = g$. So $\varphi([f]) = e$.

□

Recall the following property:

Proposition 7.41. *If X is simply connected then any two paths f, g from x_0 to x_1 are path-homotopic.*

Proof. $f * \overline{g}$ is a loop at x_0, so $f * \overline{g} \simeq_p e_{x_0}$ (X simply connected). Then $f \simeq_p f * (\overline{g} * g) \simeq_p (f * \overline{g}) * g \simeq_p e_{x_0} * g \simeq_p g$.

□

This implies the following theorem.

Theorem 7.42. *If $p : E \to B$ is a covering and E is simply connected, then $\varphi : \pi_1(B, b_0) \to p^{-1}(b_0)$ is a bijection.*

Proof. By the above, φ is surjective. If $\varphi([f]) = \varphi([g])$, then $\widetilde{f}, \widetilde{g}$ are paths in E starting at e_0 and ending at the same point e_1. Since E is simply connected, $\widetilde{f} \simeq_p \widetilde{g}$. Hence $p \circ \widetilde{f} \simeq_p p \circ \widetilde{g}$, ie. $f \simeq_p g$, so $[f] = [g]$. So φ is injective.

□

Theorem 7.43.

$$\pi_1(S^1) \simeq \mathbb{Z}$$

Proof. Consider the covering map $p : (\mathbb{R}, 0) \to (S^1, (1, 0)), p(x) = (\cos 2\pi x, \sin 2\pi x)$. Since $\mathbb{R}$ is simply connected, by the above theorem the lifting correspondence

$$\varphi : \pi_1(S^1, (1, 0)) \to p^{-1}((1, 0)) = \mathbb{Z}$$

is a bijection. We just need to show it is a group homomorphism. Let $[f], [g] \in \pi_1(S^1)$ and let $\varphi([f]) = n, \varphi([g]) = m$, ie. the lifts $\widetilde{f}$ and $\widetilde{g}$ starting at 0 ending at n and m. Define a new path $h : I \to \mathbb{R}$ by $h(s) = n + \widetilde{g}(s)$: this is the lift of

g starting at $n = \widetilde{f}(1)$. Then $\widetilde{f} * h$ is a well-defined path in $\mathbb{R}$, from 0 to $n+m$, and it is the lift of $f * g$ starting at 0. So $\varphi([f * g]) = n + m = \varphi([f]) + \varphi([g])$.

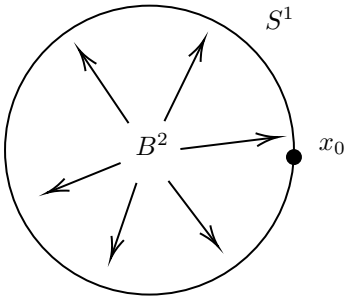
□

Remark 7.44. *We can show similarly for a torus, $\pi_1(S^1 \times S^1) \simeq \mathbb{Z} \times \mathbb{Z}$, using covering $p \times p : \mathbb{R}^2 \to S^1 \times S^1$.*

7.5 The Brouwer Fixed Point Theorem

Let B^n denote the closed ball of radius 1 in $\mathbb{R}^n$, with boundary the unit sphere S^{n-1}. Recall that, if $A \subset X$, a **retraction** $r : X \to A$ is a continuous map such that $r(a) = a \forall a \in A$.

Theorem 7.45. *There is no retraction of B^2 onto S^1.*



Proof. If $r : B^2 \to S^1$ is a retraction, then $i \circ r = \mathrm{id}_{S^1}$, so

$$\pi_1(S^1, x_0) \overset{\to}{i_*} \pi_1(B^2, x_0) \overset{\to}{r_*} \pi_1(S^1, x_0)$$

and we have $i_* \circ r_* =$ trivial homomorphism $\neq \mathrm{id} : \mathbb{Z} \to \mathbb{Z}$, contradiction.

□

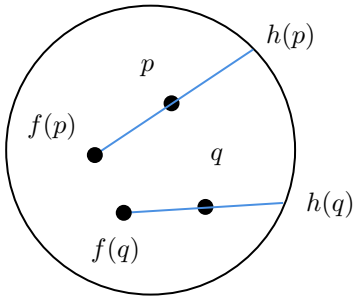
Remark 7.46. *The more elementary way to say this: given a nontrivial loop f in S^1, $i \circ f$ is nullhomotopic in B^2, via some homotopy from f to e_{x_0}. Then $r \circ H$ is a path-homotopy $r \to e_{x_0}$ in S^1, contradiction. With more algebraic topology, similarly $\nexists$ retraction $B^n \to S^{n-1} \forall n$.*

This implies the Brouwer fixed point theorem:

Theorem 7.47 (Brouwer Fixed Point Theorem)**.** *If $f : B^2 \to B^2$ is continuous, then $\exists x \in B^2$ such that $f(x) = x$.*

Remark 7.48. *With more algebraic topology, the same holds for continuous maps $B^n \to B^n \forall n$.*

Proof. Assume $f : B^2 \to B^2$ continuous, $f(x) \neq x \forall x \in B^2$. Then define $h : B^2 \to S^1$ by mapping each $p \in B^2$ to the point where the ray from $f(P)$ to P hits $\partial B^2 = S^1$.



Formula: $h(p) = p + t(p - f(p))$, where $t > 0$ such that $||h(p)||^2 = 1$. We can solve this using the quadratic formula, so t does depend continuously on p.

This gives a continuous map $h : B^2 \to S^1$, moreover if $p \in S^1$ then $h(p) = p$, so we get a retraction $B^2 \to S^1$. Contradiction.

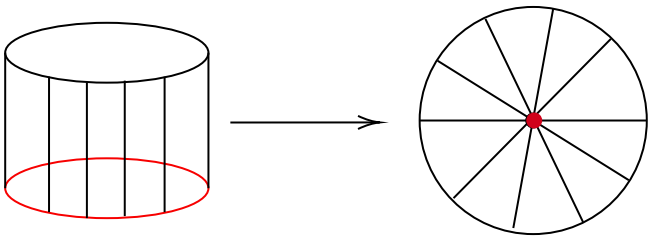
□

A loop in (X, x_0) is defined as a map $I \to X$ such that $\{0, 1\} \to \{x_0\}$, but since $I/0 \sim 1$ is homeomorphic to S^1, can also think of it as a map $(S^1, p_0) \xrightarrow{f} (X, x_0)$. So $\pi_1(X, x_0)$ tells us about homotopy classes of maps $(S^1, p_0) \to (X, x_0)$... but also $S^1 \to X$.

Lemma 7.49. *Let $h : S^1 \to X$ continuous, then the following are equivalent:*

1. *h is nullhomotopic*
2. *h extends to a continuous map $k : B^2 \to X$ $(k|_{\partial B^2 = S^1} = h)$*
3. *$h_* : \pi_1(S^1) \to \pi_1(X)$ is the trivial homomorphism.*

Proof. 1 $\implies$ 2: the key observation is that $S^1 \times I^2 \xrightarrow{p} B^2, (x, t) \mapsto t \cdot x$ is a quotient map, ie. $B^2 \simeq S^2 \times I/(x, 0) \to (x', 0) \forall x, x'$.



So: given a homotopy $H : S^2 \times I \to X$ between a constant map and $h : S' \to X$, $H(x,0) = H(x',0) \forall x, x' \in S^1$. It factors through the quotient

$$\begin{array}{ccccc} & & \overset{H}{\frown} & & \\ S^1 \times I & \xrightarrow{p} & B^2 & \xrightarrow{\exists k} & X \end{array}$$

In other terms: we can define $k : B^2 \to X$ by $k(t \cdot x) = H(x,t)$ despite angular coordinate x not being well-defined at $t = 0$, and k is continuous. So by construction $k|_{S'} = h$.

$2 \implies 3$: if $h = k|_{S'}$ then one can write $h = k \circ i$ where $i : S^1 \to B^2$ is the inclusion. By functoriality of π_1, $h_* = k_* \circ i_*$:

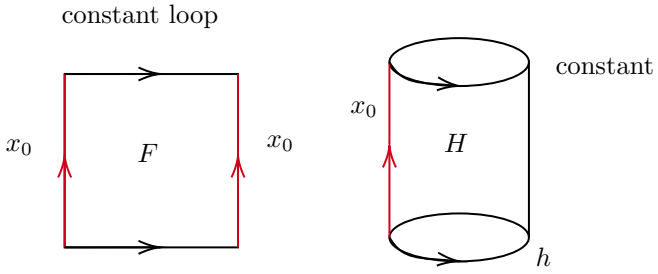
$$\begin{array}{ccccc} \pi_1(S^1) & \xrightarrow{i_*} & \pi_1(B^2) & \xrightarrow{k_*} & \pi_1(X) \\ & & \underset{h_*}{\smile} & & \end{array}$$

but $\pi_1(B^2) = \{1\}$, so h_* is trivial and so is h_*.

$3 \implies 1$: $h_* : \pi_1(S^1) \to \pi_1(X)$ trivial $\implies$ the loop $f : I \to X, s \mapsto h(e^{2\pi i s})$ $(= h \circ (\text{standard loop going around } S^1))$ represents the trivial element of $\pi_1(X, x_0)$ $(x_0 = h(1))$ hence $\exists$ path-homotopy $F : I \times I \to X$ from f to constant loop at x_0; note that $F(0,t) = F(1,t) = x_0 \forall t \in I$. Recall $I \times I/(0,t) \sim (1,t) \forall t$ is homeomorphic to $S^1 \times I$. This implies F factors through the quotient:

$$\begin{array}{ccccc} & & \overset{F}{\frown} & & \\ I \times I & \xrightarrow[p]{} & S^1 \times I & \xrightarrow{\exists H} & X \end{array}$$

H gives a homotopy from h to constant map.



□

Exercise 7.50. *The inclusion $S^1 \hookrightarrow \mathbb{R}^2 - \{0\}$ and the identity map $S^1 \to S^1$ aren't nullhomotopic, using lemma and i_* nontrivial on π_1.*

Let's look at another application: the fundamental theorem of algebra.

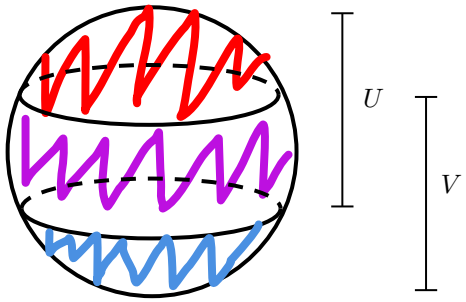
Theorem 7.51 (The Fundamental Theorem of Algebra)**.** $f(z) = z^d + a_{d-1}z^{d-1} + \ldots + a_0$ *complex polynomial of deg* $d > 0 \implies \exists z_0 \in \mathbb{C}$ *such that* $f(z_0) = 0$.

Proof. For $|z| = r > 0$ the term z^d dominates (as soon as $r^k > d|a_{d-k}| \forall 1 \leq k \leq d$) so that $|a_{d-k}z^{d-k}| < \frac{1}{d}r^d$, so straight line segment $f(z) \to z^d$ doesn't cross 0. This implies $F(z,t) = (1-t)f(z) + tz^d$ has no zeroes on $\{|z| = r\} \times I$. Hence: the maps $S^1 \to S^1$ defined by $e^{i\theta} \mapsto \frac{f(re^{i\theta})}{|f(re^{i\theta})|}$ and $e^{i\theta} \mapsto e^{ni\theta}$ are homotopic via $(e^{i\theta}, t) \mapsto F(re^{i\theta}, t)/|F(re^{i\theta}, t)|$. These are nontrival on $\pi_1(S^1)$ (in fact, map generator $1 \in \mathbb{Z}$ to $d \in \mathbb{Z}_{>0}$) hence don't extend over B^2. But if f had no roots, $z \mapsto \frac{f(rz)}{|f(rz)|}$ would be such an extension.

□

Now, we'll provide a short introduction to the Seifert-Van Kampen Theorem.

Question: Assume $X = U \subset V$, with U and V open subsets, and we know $\pi_1(U)$ and $\pi_1(V)$. Can we find $\pi_1(X)$? Eg.



$S^2 = U \cup V, \pi_1(U)$ and $\pi(V)$ trivial, then

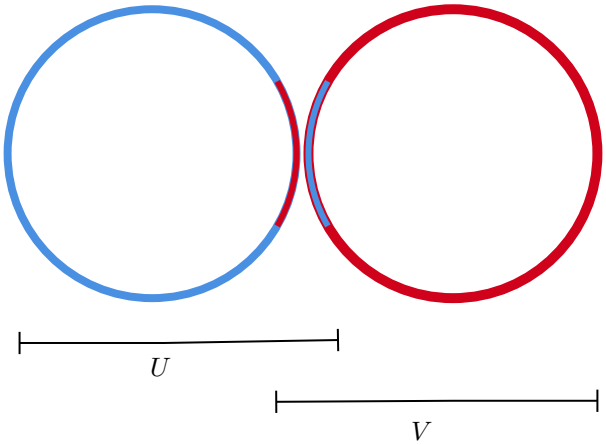


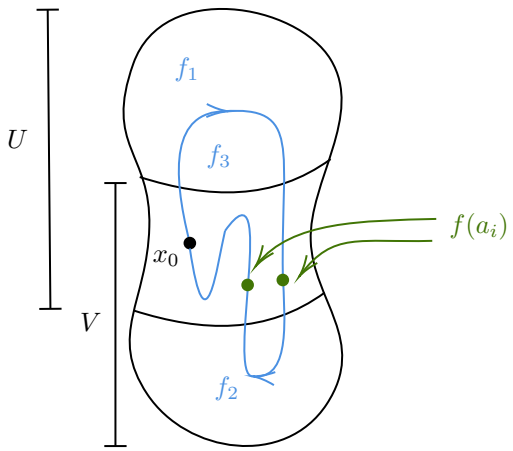
figure 8 $= U \cup V$, each of U and V has homotopy type of S^1.

The Seifert-Van Kampen Theorem, which we'll see soon, gives a general way to calculate $\pi_1(X)$ in this situation. For now, we'll just prove a weaker (and easier version).

Theorem 7.52. *Suppose $X = U \cup V$, U and V open, $U \cap V$ path-connected, $x_0 \in U \cap V$. Let $i : U \hookrightarrow X$ and $j : V \hookrightarrow X$ be the inclusion maps. Then the images of $i_* : \pi_1(U, x_0) \to \pi_1(X, x_0)$ and $j_* : \pi_1(V, x_0) \to \pi_1(X, x_0)$ generate $\pi_1(X, x_0)$.*

Ie: every element of $\pi_1(X, x_0)$ can be expressed as a product of elements in Im (i_*) and Im (j_*), or every loop in (X, x_0) is path-homotopic to a composition of loops entirely contained in either U or V.

Proof. Let $f : I \to X$ be a loop based at x_0. $[0, 1] = f^{-1}(U) \cup f^{-1}(V)$ open cover, $[0, 1]$ compact.

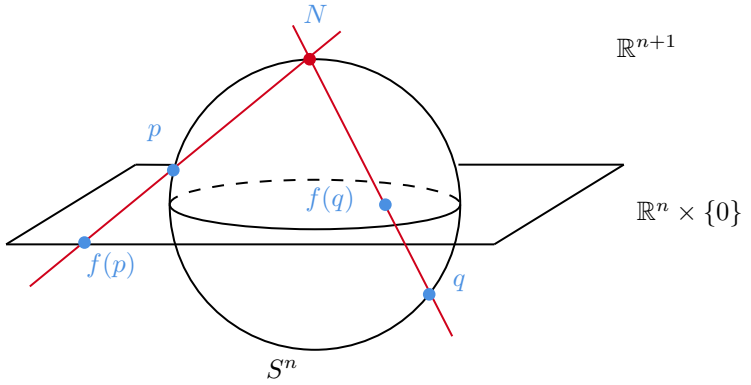


using the Lebesgue number lemma, we can subdivide $[0, 1]$ into $0 = a_0 < a_1 < \ldots < a_n = 1$ such that $f([a_{i-1}, a_i])$ is contained in either U or V. Eliminating unnecessary a_i from the list, we can assume U and V alternate along the way, and in particular $f(a_i) \in U \cap V \forall i$. Let $f_i = f|_{[a_{i-1}, a_i]}$ so that $[f] = [f_1] * \ldots * [f_n]$. For each i, choose path α_i in $U \cap V$ from x_0 to $f(a_i)$ (take $\alpha_0 = \alpha_n =$ constant path at x_0). Then $[f] = [\alpha_0 * f_1 * \alpha_1^{-1}] * [\alpha_1 * f_2 * \alpha_2^{-1}] * \ldots * [\alpha_{n-1} * f_n * \alpha_n^{-1}]$, and we are done.

□

Corollary 7.53. *$X = U \cup V$ with U and V open and simply-connected, $U \cap V$ path-connected $\implies$ X is simple-connected.*

Example 7.54. *Let $X = S^n, n \geq 2$, and $U = S^n - (0,0,...,0,1), V = S^n - (0,...,0,-1)$. Then U and V are homeomorphic to $\mathbb{R}^n$ via* ***stereographic projection*** *$f : U \to \mathbb{R}^n$ mapping each point $x \in U$ to the point where the line in $\mathbb{R}^{n+1}$ through N and x intersects the equatorial plane $\mathbb{R}^n \times \{0\}$*



ie. $f(x) = \frac{1}{1-x_{n+1}}(x_1,...,x_n)$ (change $-$ to $+$ for $V \xrightarrow{\sim} \mathbb{R}^n$). Hence: U and V, homeomorphic to $\mathbb{R}^n$ are simply connected and $U \cap V$ homeomorphic to $\mathbb{R}^n - \{point\}$ is path-connected ($n \geq 2$)

Corollary 7.55. *S^n is simply connected for $n \geq 2$.*

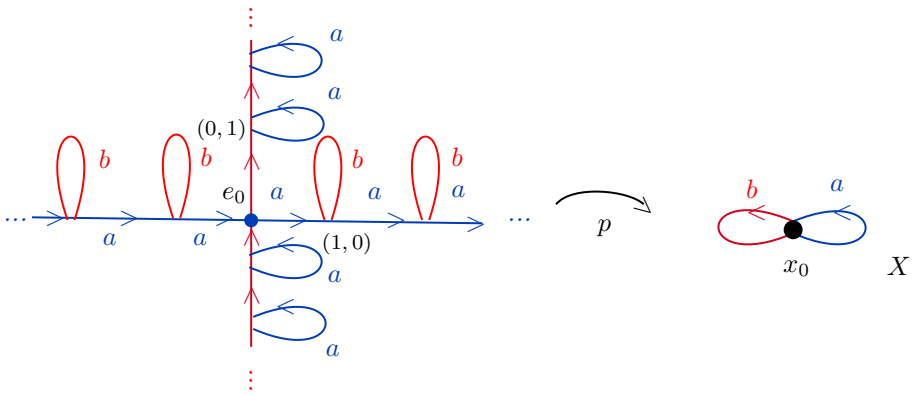
Corollary 7.56. *An open subset in $\mathbb{R}^n \forall n \geq 3$ cannot be homeomorphic to an open subset in $\mathbb{R}^2$.*

Indeed: $U \subset \mathbb{R}^n$ open, $p \in U \implies \exists$ open ball $p \in B_r(p) \subset U$, and $B_r(p) - \{p\}$ deformation retracts onto a sphere $\implies$ $B_r(p) - \{p\}$ is simply connected. Whereas $q \in V \subset \mathbb{R}^2$ open $\implies \forall$ open $q \in N \subset V, N - \{q\}$ can't be simply connected (retracts to circle). The same argument for $\mathbb{R}^n$ for $n \geq 2$ vs $\mathbb{R}$ is easier, only uses connectedness.

Example 7.57. *Recall from HW that the quotient of S^n by $x \sim -x, p : S^n \to S^n/\sim \simeq \mathbb{RP}^n$ is a degree 2 covering map. Also recall: lifting correspondence $\pi_1(\mathbb{RP}^n, b_0) \to p^{-1}(b_0) = \{2 \text{ points}\}$ surjective because S^n connected; injective because S^n is simply connected if $n \geq 2$. (If a loop in $\mathbb{RP}^n$ lifts to a loop $\widetilde{f}$ in S^n, then $\widetilde{f}$ is homotopic to constant loop in S^n, and projecting by p, $p \circ \widetilde{f} = f$ is homotopic to a constant loop in $\mathbb{RP}^n$.) For $n \geq 2, \pi_1(\mathbb{RP}^n)$ is a group with 2 elements, hence isomorphic to $\mathbb{Z}/2\mathbb{Z}$.*

Example 7.58. *Let X be the figure 8 space oriented counterclockwise. We can cover this by open sets U, V which have deformation retractions to $S^1, U \cap V$ connected. By theorem, $\pi_1(X)$ is generated by the images of two maps from $\mathbb{Z}$, ie. can express every loop in terms of powers of $[a]$ and $[b]$ (a, b loops around each S^1) generators of $\pi_1(U), \pi_1(V)$, ie. every element is a product of $[a]^{\pm 1}$'s and $[b]^{\pm 1}$'s. But we don't know the relations between $[a]$ and $[b]$. We can show that $[a]$ and $[b]$ don't commute - $[a] * [b] \neq [b] * [a]$. One way to do this is by*

looking at covering map



The lift of $a*b$ starting at e_0 ends at $(1,0)$ and the lift of $b*a$ starting at e_0 ends at $(0,1)$, hence $[a]*[b] \neq [b]*[a]$ so $\pi_1(X, x_0)$ is not abelian. In fact, we'll show later that it is the free group generated by $[a]$ and $[b]$, ie. elements are arbitrary words in $[a]^{\pm 1}$ and $[b]^{\pm 1}$ with no relations whatsoever (except $[a]^{-1} * [a] = 1$, etc.)

7.6 Equivalence and More About Covering Spaces

Question: Let $p : (E, e_0) \to (B, b_0)$ covering map. How are $\pi_1(E)$ and $\pi_1(B)$ related? (Always assume E and B are path connected).

Theorem 7.59. *$p_* : \pi_1(E, e_0) \to \pi_1(B, b_0)$ is an injective homomorphism.*

Proof. If $\widetilde{h}$ is a loop at e_0 and $p_*([\widetilde{h}]) = \mathrm{id}$, then $\exists$ path-homotopy $H : I \times I \to B$ from $p \circ \widetilde{h}$ to the constant loop at b_0. Its lift $\widetilde{H}$: $I \times I \to E$ starting at e_0 is then a path-homotopy from $\widetilde{h}$ to the constant loop, so $[\widetilde{h}] = \mathrm{id}$.

□

Hence, the covering $p : E \to B$ gives a subgroup $H : \mathrm{Im}\ (p_*) = \pi_1(B, b_0)$, with

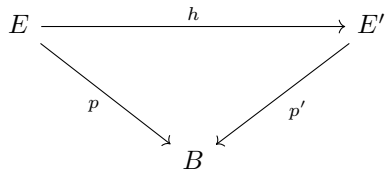
$$\pi_1(E, e_0) \underset{p_*}{\overset{\text{iso}}{\widetilde{\to}}} H.$$

It turns out that:

1. The subgroup $H \subset \pi_1(B, b_0)$ determines the covering p.
2. Assuming B is path-connected and "sufficiently nice" ("semi-locally simply connected"), for each subgroup H of $\pi_1(B, b_0), \exists$ covering $p : E \to B$ such that $p_*(\pi_1(E)) = H$.

Now let's discuss equivalence of covering spaces.

Definition 7.60. *Let $p : E \to B, p' : E' \to B$ coverings. p and p' are **equivalent** if $\exists$ homeomorphism $h : E \to E'$ such that $p = p' \circ h$:*



*Then we say h is an **equivalence of coverings**.*

$\forall b \in B, h$ gives a bijection $p^{-1}(b) \xrightarrow{\sim} p'^{-1}(b)$ between the sheets of p and p'. By continuity, over a connected evenly covered subset $U \subset B$ this looks like $p^{-1}(U) \simeq U \times a \xrightarrow[\mathrm{id}\times\sigma]{} U \times A' \simeq p'^{-1}(U)$, where $\sigma : A \to A'$ is a bijection between sets of sheets.

Goal: if two coverings have same corresponding subgroups of $\pi_1(B)$ then they are equivalent. For this we need a general lifting lemma.

Definition 7.61. *A space X is locally path-connected if $\forall x \in X, \forall U \ni x, \exists V \subset U$ path-connected neighborhood of x.*

Counterexample: $(\{\frac{1}{n}, n \geq 1\} \cup \{0\}) \times \mathbb{R} \cup \mathbb{R} \times 0$ in $\mathbb{R}^2$ is path-connected but not locally path-connected.

From now on, assume $p : E \to B$ covering, E and B path-connected and locally path-connected.

Theorem 7.62 (Lifting Lemma for Loops)**.** *A loop f in (B, b_0) lifts to a loop in (E, e_0) if and only if $[f] \in p_*(\pi_1(E, e_0)) \subset \pi_1(B, b_0)$.*

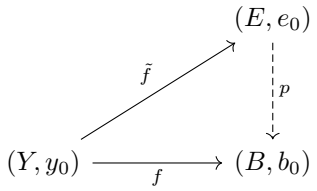
Proof. If f can be lifted to $\widetilde{f}$ of f at e_0 is a loop in E, then $[f] = [p \circ \widetilde{f}] = p_*([\widetilde{f}]) \in p_*(\pi_1(E))$.

If $[f] = p_*([\widetilde{g}])$ for some loop $\widetilde{g}$ in (E, e_0) then $p \circ \widetilde{g}$ is path-homotopic to f. Lifting this path-homotopy to E, we get a path-homotopy in E between $\widetilde{g}$ and the lift $\widetilde{f}$ of f. Since $\widetilde{g}$ is a loop, so is $\widetilde{f}$.

□

Theorem 7.63 (General Lifting Lemma)**.** *Let $p : E \to B$ covering map, $p(e_0) = b_0$. Let Y be path-connected and locally path-connected, and $f : Y \to B$ continuous map such that $f(y_0) = b_0$. Then f can be lifted to $\widetilde{f} : Y \to E$ with $\widetilde{f}(y_0) = e_0$ if and only if $f_*(\pi_1(Y, y_0)) \subset p_*(\pi_1(E, e_0))$. If it exists, the lift is*

unique:



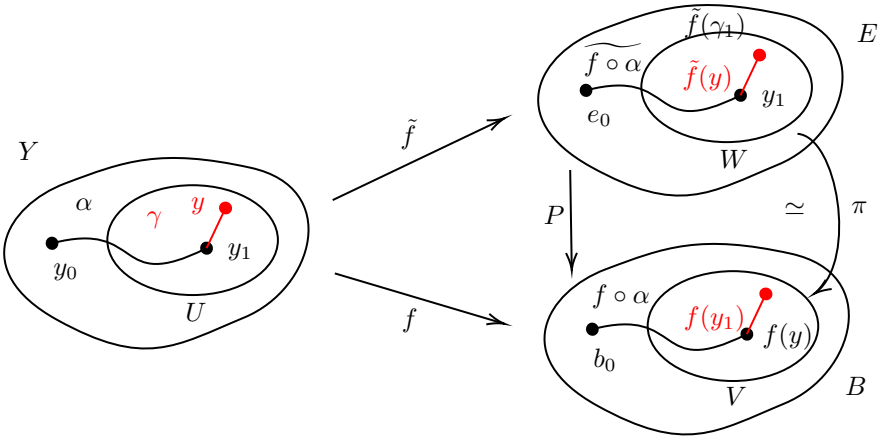
Proof. If f can be lifted to $\widetilde{f}$, then $f = p \circ \widetilde{f}$ so

$$f_*(\pi_1(Y, y_0)) = p_*(\widetilde{f}_* (\pi_1(Y, y_0))) \subset p_*(\pi_1(E, e_0)).$$

Conversely, assume the condition holds, and let $y_1 \in Y$. Choose a path α from y_0 to y_1 in Y. Lift $f \circ \alpha : I \to B$ to a path in E starting at e_0. Define $\widetilde{f}(y_1) =$ the end point of this path. (this is the only possibility for $\widetilde{f}(y_1)$ if a continuous lift exists, since the unique lift of $f \circ \alpha$ will then be $\widetilde{f} \circ \alpha$). It remains to check that $\widetilde{f}$ is well-defined and continuous.

Well-defined. Let β be a different path in Y from y_0 to y_1. Then $\alpha * \overline{\beta}$ is a loop in (Y, y_0), $f \circ (\alpha * \overline{\beta})$ loop in (B, b_0), representing $f_*[\alpha * \overline{\beta}] \in \text{Im } f_* \subset p_*(\pi_1(E, e_0))$ so it lifts to a loop in E (by previous theorem). So $f \circ \alpha$ lifts to a path from e_0 to $\widetilde{f}(y_1)$ as defined above, and $\overline{f \circ \widetilde{\beta}}$ lifts to a path from $\widetilde{f}(y_1)$ back to e_0, hence $f \circ \beta$ lifts to a path from e_0 to $\widetilde{f}(y_1)$. Thus $\widetilde{f}(y_1)$ is independent of the choice of path $y_0 \to y_1$.

Continuity of $\widetilde{f}$: enough to check on a neighborhood of y_1. Let $V \subset B$ be an evenly covered neighborhood of $f(y_1)$, and using a local path-connectedness of Y, can find $U \subset f^{-1}(V)$ path-connected neighborhood of y_1 in Y. Let $W \subset p^{-1}(V) \subset E$ be the slice containing $\widetilde{f}(y_1); p|_W = \pi : W \xrightarrow{\sim} V$ homeomorphic.



For $y \in Y, \exists$ path γ in U from y_1 to y, and $\pi^{-1} \circ f \circ \gamma$ is a lift of $f \circ \gamma$ to $W \subset E$ starting at $\widetilde{f}(y_1)$. And so the lift of $f \circ (\alpha * \gamma)$ to E starting at e_0 is composition of $f \widetilde{\circ} \alpha$ (from e_0 to $\widetilde{f}(y_1)$) and $\pi^{-1} \circ f \circ \gamma$ from $\widetilde{f}(y_1) = \pi^{-1}(f(y_1))$ to $\pi^{-1}(f(y))$. Hence $\widetilde{f}(y) = \pi^{-1}(f(y))$. So $\widetilde{f}|_U = \pi^{-1} \circ f|_U$ is continuous, and hence $\widetilde{f}$ is continuous.

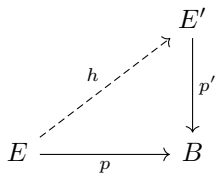
□

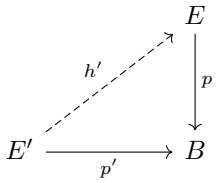
Now we can tell when two coverings are equivalent, as long as all maps preserve base points!

Theorem 7.64. *Let $p : E \to B, p' : E' \to B$ covering maps with $p(e_0) = p'(e_0') = b_0$. There is an equivalence $h : E \xrightarrow{\sim} E'$ such that $h(e_0) = e_0'$ if and only if the subgroups $H = p_*(\pi_1(E, e_0))$ and $H' = p'_*(\pi_1(E', e_0'))$ are equal (the same subgroup of $\pi_1(B, b_0)$). Moreover, if h exists it is unique.*

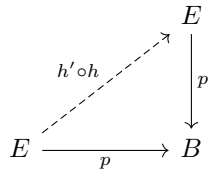
Proof. $\Longrightarrow$: if $H : E \to E'$ is an equivalence with $h(e_0) = e_0'$, then $h_*(\pi_1(E, e_0)) = \pi_1(E', e_0')$. The conclusion then follows from $p'_* \circ h_* = p_*$.

$\Longleftarrow$: assume $H = H'$. Then by lifting lemma, $\exists$ unique base point preserving lifts





so $p' \circ h = p$ and $p \circ h' = p'$. Now $p \circ h' \circ h = p' \circ h = p$, so $h' \circ h : E \to E$ is a lifting



But so is id_E. By uniqueness of lifting, we get $h' \circ h = \mathrm{id}_E$. Similarly $h \circ h' = \mathrm{id}_{E'}$. So h is a homeomorphism such that $p' \circ h = p$, hence an equivalence of coverings.

□

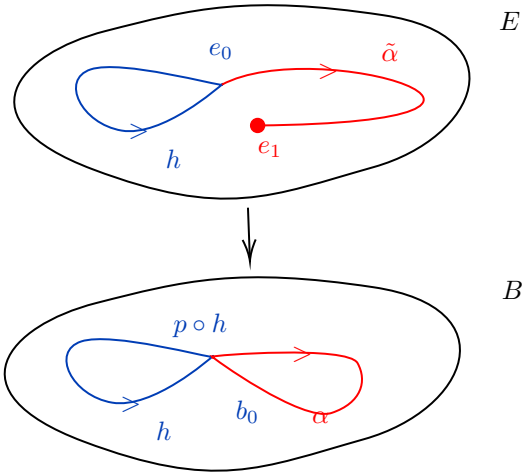
Example 7.65. *These are all subgroups of $\mathbb{Z}$, so every connected covering of S^1 is equivalent to exactly one of these.*

- $p_k : S^1 \to S^1, z \mapsto z^k$, $(p_k)_* : \pi_1(S^1, b_0) \to \pi_1(S^1, b_0)$ *multiplication by* $k \implies H = k\mathbb{Z} \subset \mathbb{Z}$.
- $p_0 : \mathbb{R} \to S^1, x \mapsto (\cos x, \sin x)$, $(p_0)_*(\pi_1(\mathbb{R})) = \{0\}$.

What if we consider equivalence $h : E \to E'$ that don't map e_0 to e_0'?

Then the corresponding subgroups of $\pi_1(B, b_0)$ are conjugate.

Indeed, if we change the base point in a (path-connected) covering spae $p : E \to B$... if $e_0, e_1 \in p^{-1}(b_0)$ and $\widetilde{\alpha}$ is a path from e_0 to e_1, recall $\pi_1(E, e_0) \xrightarrow{\sim} \pi_1(E, e_1), [h] \mapsto [\overline{\alpha}^{-1} * h * \widetilde{\alpha}]$. Then $\alpha = p \circ \widetilde{\alpha}$ is a loop in (B, b_0), so whenever $[p \circ h] = p_*([h]) \in H_0 = p_*(\pi_1(E, e_0)) \implies [\alpha]^{-1} * [p \circ h] * [\alpha] \in H_1 = p_*(\pi_1(E, e_1))$.



So $[\alpha]^{-1}H_0[\alpha] \subset H_1$ and similarly in the reverse direction $[\alpha]H_1[\alpha]^{-1} \subset H_0$, hence equal.

Conversely, if H_0, H_1 are conjugate subgroups of $\pi_1(B, b_0)$, ie. $\exists[\alpha]$ such that $H_1 = [\alpha]^{-1}H_U[\alpha]$ and $H_0 = p_*(\pi_1(E, e_0))$ then let $\widetilde{\alpha}$= lift of α to a path in E starting at e_0, and let $e_1 = \widetilde{\alpha}(1)$, then $H_1 = p_*(\pi_1(E, e_1))$.

This implies the following theorem:

Theorem 7.66. *$p : E \to B, p' : E' \to B$ covering maps, $p(e_0) = p'(e_0') = b_0$. Then p and p' are equivalent $\iff$ the subgroups $H = p_*(\pi_1(E, e_0)), H' = p_*'(\pi_1(E', e_0'))$ of $\pi_1(B, b_0)$ are conjugate.*

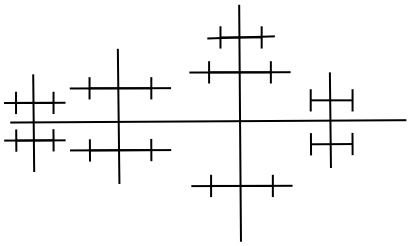
7.7 Universal Enveloping Space

Definition 7.67. *Let $p_0 : E_0 \to B$ covering and E_0 is simply connected, say E_0 is a **universal covering** of B.*

Note: this corresponds to the trivial subgroup $p_{0_*}(\pi_1(E_0)) = \{1\} \subset \pi_1(B)$, unique up to equivalence by the above.

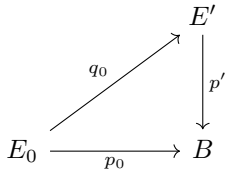
Example 7.68.

- $p : \mathbb{R} \to S^1, p \times p : \mathbb{R}^2 \to S^1 \times S^1 =$ *torus*
- *Infinite tree $\to \infty$, (horizontal edges $\to a$, vertical edges $\to b$)*



Theorem 7.69. *Let $p_0 : E_0 \to B$ universal covering, $p' : E' \to B$ any path-connected covering, then $\exists$ covering map $q_0 : E_0 \to E'$ such that $p' \circ q_0 = p_0$ and q_0 is the universal covering of E'*

q_0 is constructed by lifting:



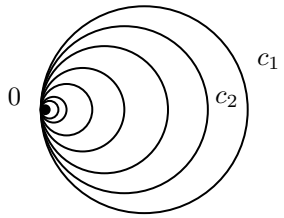
$\exists$ since $p_{0_*}(\pi_1(E_0)) = \{1\} \subset p'_*(\pi_1(E'))$ and can show that it's a covering map as well.

So in fact, if B has a universal covering, all other coverings can then be obtained as quotients.

Some spaces have no universal coverings!

Example 7.70.

- *Hawaiian earrings* $= \bigcup_{n\geq 1} C_n$ *circles of radius* $\frac{1}{n}$ *called* $\left(\frac{1}{n}, 0\right)$ *inside* $\mathbb{R}^2$.



Any covering space must evenly over a neighborhood of the origin, which presents it from being simply connected. (for n sufficiently large, loop around C_n lifts to a loop).

If one avoids such pathological examples - assuming B is (semi) locally simply connected, can build universal cover as space of pairs (b, γ) where $b \in B, \gamma =$ homotopy class of path $b_0 \to b$.

This has a preferred topology for which any simply connected neighborhood $U \ni b$ is evenly covered: if $b' \in U$, adding a path $b \to b'$ inside U on its

inverse gives a preferred bijection {homotopy classes of paths $b_0 \to b$} $\iff$ {homotopy classes of paths $b_0 \to b'$} independent of choice of path $b \to b'$ inside U since U is simply connected.

7.8 Free Products

Given $X = U \cup V$, $U, V, U \cap V \subset X$ open and path connected, this describes $\pi_1(X)$ in terms of $\pi_1(U)$ and $\pi_1(V)$. We've already seen a simple statement: $\pi(X)$ is generated by the images of $\pi_1(U) \xrightarrow{i_*} \pi_1(X), \pi_1(V) \xrightarrow{j_*} \pi_1(X)$

To formulate the theorem, we need to discuss the notion of free product of groups.

Assume G is a group, $G_1, ..., G_n$ subgroups of G which generate G, ie. any $x \in G$ can be written as $x = x_1...x_m$ where each x_i is in some G_j. Also assume $G_j \cap G_k = \{1\} \forall j \neq k$, $(x_1, ..., x_m)$ is called a **word** of length m that represents x.

Say $(x_1...x_m)$ is **reduced word** if no G_j contains two consecutive elements x_i, x_{i+1}. (in particular if $m \geq 2$, no x_i can be $= 1$). (else can reduce to a shorter word $(x_1, ..., x_i, x_{i+1}, ..., x_m)$).

Definition 7.71. *G is the **free product** of the subgroups $G_1, ..., G_n$, denoted $G = G_1 * ... * G_n$ if G_i generate G, $G_i \cap G_j = \{1\}$, and every element of G is represented by a unique reduced word.*

Example 7.72. *$\mathbb{Z}^2$ is not the free product of its two factors: denoting by a and b the two generators ($a = (1,0), b = (0,1)$), $ab = ba$ is represented by reduced words $(a, b), (b, a), (a^2, b, a^{-1}), ...$*

Alternative characterization (universal property): G is the free product of subgroup G_j's iff, for any group H and any homomorphisms $h_j : G_j \to H, \exists$ unique homomorphism $h : G \to H$ such that

$$G_j \hookrightarrow G \xrightarrow{h} H, \quad G_j \xrightarrow{h_j} H$$

commutes $\forall j$.

The point is: uniqueness of expression allows us to defined $h(x_1, ..., x_m) = h_{j_1}(x_i)...h_{j_m}(x_m)$.

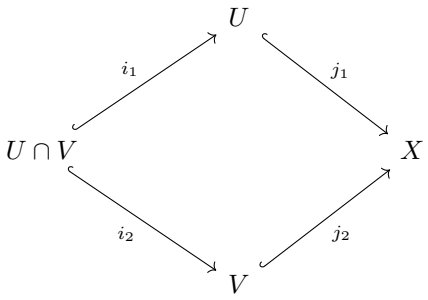
Definition 7.73. *The **external free product** of groups is the group G plus injective homomorphisms $G_j \xrightarrow{ij} G$ such that G is the free product of the subgroups $i_j(G_j)$.*

Proposition 7.74. *The external free product always exists and is unique up to isomorphism. It can be constructed as set of reduced words in G_j's (with product = concatenate and reduce) and satisfies universal property.*

In particular the **free group** on the elements $\{a_j\}$ is defined to be the external free product of cyclic groups $G_j = \{a_j^n | n \in \mathbb{Z}\} (\simeq \mathbb{Z})$.

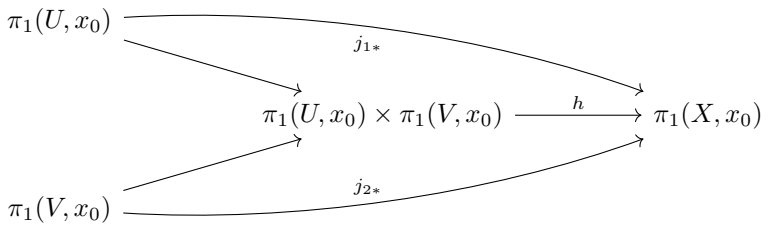
7.9 Seifert-Van Kampen

Let $X = U \cup V$, U and V open in X, $U \cap V$ path-connected $\ni x_0$. The inclusions



induce homomorphisms of π_1.

By the universal property of free products, $\exists$ unique homomorphism h such that



commutes.

(Define h on words in elements of $\pi_1(U, x_0)$ and $\pi_1(V, x_0)$ using j_{1*} and j_{2*} on each component of the word!)

Theorem 7.75 (Seifert-Van Kampen)**.** *The homomorphism h defined above is surjective, and its kernel N is the smallest normal subgroup of $\pi_1(U, x_0) * \pi_1(V, x_0)$ which contains all elements of the form $i_{1*}(g)^{-1} * i_{2*}(g) \forall \pi_1(U \cap V, x_0)$, ie. $\pi_1(X, x_0) \cong \pi_1(U, x_0) * \pi_1(V, x_0)/N$.*

Corollary 7.76. *If $U \cap V$ is simply connected then $\pi_1(X, x_0) \cong \pi_1(U, x_0) * \pi_1(V, x_0)$*

Corollary 7.77. *If V is simply connected then $\pi_1(X, x_0) \cong \pi_1(U, x_0)/N$ where N is the smallest normal subgroup containing the image of $i_{1*} : \pi_1(U \cap V, x_0) \to \pi_1(U, x_0)$.*

Example 7.78. *Figure 8 $\implies$ U, V deformation retract onto circles, $U \cap V$ contractible. Hence $\pi_1(X, x_0) \cong \pi_1(U, x_0) * \pi_1(V, x_0) \cong \mathbb{Z} \times \mathbb{Z}$ free group generated by loops around the two circles.*

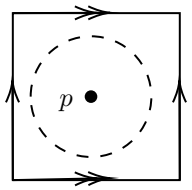
Example 7.79. *By induction, wedge of n circles:*

$X = \bigcup_{i=1}^{n} S_i$, S_i homeomorphic to $S^1 \forall i, S_i \cap S_j = \{x_0\} \implies \pi_1(X, x_0) =$ free group on n generators $a_i =$ loops generating $\pi_1(S_i, x_0)$. (Similarly for a finite graph with n loops).

7.10 Fundamental Groups of Surfaces

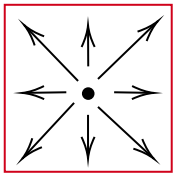
We can also calculate these using Van Kampen!

Example 7.80. *Let's calculate π_1 of a torus (easiest is still $\mathbb{R}^2 \xrightarrow[p \times p]{} T$ for an universal cover T). $T \simeq I \times I/(x,0) \sim (x,1) \forall x, (0,y) \sim (1,y) \forall y$.*



Let $U = T \backslash \{p\}$, $V =$ open ball of radius $< \frac{1}{2}$ around p.

U deformation retracts onto wedge of two circles



$/ \sim$

, V is simply connected, and $U \cap V \simeq D^2$-point has boundary type of S^1.

Using corollary 2 above: $\pi_1(T) \simeq \pi_1(U)/N$, where N is the normal subgroup generated by the image of the loop f which generates $\pi_1(U \cap V)$ (and its conjugates).

$\pi_1(U)$ is a free group on generators a, b; and then the image of $[f]$ under the inclusion $U \cap V \hookrightarrow U$ is $aba^{-1}b^{-1}$. So we set $aba^{-1}b^{-1} = 1$, ie. $ab = (aba^{-1}b^{-1})ba = ba$, get abelian group $\simeq \mathbb{Z}^2$. So

$$\pi_1(T) \cong \langle a, b | ab = ba \rangle \cong \mathbb{Z} \times \mathbb{Z}.$$

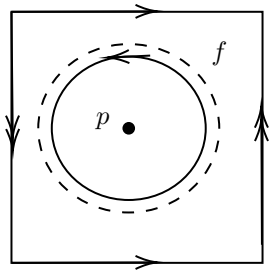
Example 7.81. *Similarly for $\pi_1(\mathbb{RP}^2)$, using $\mathbb{RP}^2 \simeq S^2/x \sim -x$, sphere $\simeq B^2/\sim$ with $x \sim -x \forall x \in S^1 = \partial B^2$. Now write $\mathbb{RP}^2 = U \cup V, U = \mathbb{RP}^2 - \{p\}, V =$ disc centered at p.*

U deformation retracts onto the boundary $S^1/x \sim -x \xrightarrow[z \mapsto z^2]{\sim} S^1$ so $\pi_1(U) \cong \mathbb{Z}$ with generator c. V is simply connected: $U \cap V \cong D^2$-point has homotopy type of S^1.

$\pi_1(\mathbb{RP}^2) \cong \pi_1(U)/N$, *N normal subgroup generated by image of generator* $[f] \in \pi_1(U \cap V)$ *under inclusion, which is* c^2. *So*

$$\pi_1(\mathbb{RP}^2) = \langle c | c^2 = 1 \rangle \simeq \mathbb{Z}/2\mathbb{Z}.$$

Example 7.82. *Klein bottle: recall* $K = I \times I/\sim, (x,0) \sim (x,1), (0,y) \sim (1, 1-y)$.



Again write $K = U \cup V, U = K - \{p\}, V =$ *disc centered at* $p \implies \pi_1(K) \cong \pi_1(U)/N$.

U *retracts on the boundary* $\cong$ *figure 8 space so* $\pi_1(U) \simeq$ *free group on generators* a, b.

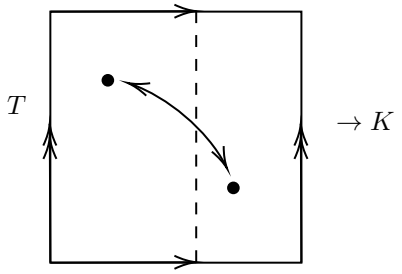
$U \cap V$ *has homotopy type of* S^1, *and the generator* $[f] \in \pi_1(U \cap V) \cong \mathbb{Z}$ *maps under inclusion to* $aba^{-1}b$.

So $\pi_1(K) \cong \langle a, b | aba^{-1}b = 1 \rangle$ *which is not abelian* ($ab = b^{-1}a$, *not* ba) *ie* $aba^{-1} = b^{-1}$ *so* b *is conjugate to its inverse.*

But this contains an index 2 subgroup H *generated by* a^z *and* b, *which commute!* ($aba^{-1} = b^{-1} \implies$ *taking inverses,* $ab^{-1}a^{-1} = b$, *so* $a^2ba^{-2} = a(aba^{-1})a^{-1}ab^{-1}a^{-1} = b \implies a^2b = ba^2$ *so the subgroup* $H \cong \mathbb{Z}^2$).

We can show, by rearranging letters via $ab = b^{-1}a$, *this contains all words with even number of* a*'s so it is an index 2 subgroup.*

This subgroup corresponds to a degree 2 covering map by the torus, $T \to K$!



Ie. map $(x,y) \in I \times I/\sim_T$ *to* $(2x, y)$ *if* $x \leq \frac{1}{2}$ *and* $(2x-1, 1-y)$ *if* $x \geq \frac{1}{2}$ *in* $I \times I/\sim_K$.

Remark 7.83. *Cool fact that this relates to: if you coat a Klein bottle in paint all over, the paint forms a torus.*

8 Real Analysis

8.1 Review: Real Functions

Recall that the basic object of real analysis are functions $f : \mathbb{R} \to \mathbb{R}$ (or a subset of $\mathbb{R}$, the domain of f) and their continuity, differentiability, integrals... plus sequences and series of functions:

Definition 8.1. *We say that a function f is **continuous at a point** x if*

$$\forall \epsilon > 0, \exists \delta > 0 \text{ such that for all } y,\ |x - y| < \delta \implies |f(x) - f(y)| < \epsilon.$$

This is equivalent to

$$\lim_{t \to x} f(t) = f(x).$$

More general limits include:

$$\lim_{x \to \infty} f(x), \quad \lim_{t \to x, t < x} f(t), \dots$$

Infinite limits and limits at infinity can be understood as taking place in the compactification $\mathbb{R} \cup \{\pm\infty\}$. For example,

$$\lim_{x \to 0, x > 0} f(x) = \infty$$

means

$$\forall M > 0, \exists \delta > 0 \text{ such that for all } x,\ 0 < x < \delta \implies f(x) > M.$$

Using compactness and connectedness of $[a, b] \subset \mathbb{R}$, we've already seen:

Definition 8.2. *A function $f : [a, b] \to \mathbb{R}$ is **uniformly continuous** if*

$$\forall \epsilon > 0, \exists \delta > 0 \text{ such that for all } x, y \in [a, b],\ |x - y| < \delta \implies |f(x) - f(y)| < \epsilon.$$

This means that the same δ works for all $x \in [a, b]$.

Theorem 8.3 (The Intermediate Value Theorem)**.** *If $f([a, b])$ is connected, then it contains all reals between $f(a)$ and $f(b)$*

Theorem 8.4 (The Extreme Value Theorem)**.** *If $f([a, b])$ is compact, then it is bounded and contains its inf and sup.*

We have considered two topologies on **spaces of functions** so far (e.g., for $\mathbb{R} \to \mathbb{R}$, and similarly for $\mathbb{R}^n \to \mathbb{R}^m$):

- **Pointwise topology:** A sequence of functions $f_n \to f$ pointwise if for every $x \in \mathbb{R}$, $f_n(x) \to f(x)$.
- **Uniform topology:** A sequence of functions $f_n \to f$ uniformly if $\|f_n - f\|_\infty := \sup_x |f_n(x) - f(x)| \to 0$.

We have also seen that if f_n is continuous and $f_n \to f$ uniformly, then f is continuous. Furthermore, spaces of functions such as $\mathbb{R} \to \mathbb{R}$ or $[a, b] \to \mathbb{R}$ (or $\mathbb{R}^n \to \mathbb{R}^m$) with the uniform topology are complete metric spaces. The space of continuous functions, C^0, is a closed subspace and hence complete as well. However, unless we restrict to bounded functions, $\sup |f - g|$ does not quite form a metric.

Analysis often involves various spaces of functions (e.g., bounded, integrable, continuous, differentiable) and different topologies (often, but not always, metrics) defined on them.

Beyond polynomials and a few other explicit examples, many functions are defined as limits of sequences or series. A key example (also relevant for complex analysis) is the **power series**, which takes the form

$$f(x) = \sum_{n=0}^{\infty} a_n x^n,$$

where $a_n \in \mathbb{R}$ are coefficients. (Simply writing this expression does not guarantee that the series converges for any $x \neq 0$.) We will need to understand convergence (pointwise, uniformly over certain subsets of $\mathbb{R}$, etc.), so basic facts about real sequences and series in $\mathbb{R}$ will come in handy.

8.2 Review: Sequences and Series in $\mathbb{R}$

Since $\mathbb{R}$ is complete:

Proposition 8.5. *A sequence in $\mathbb{R}$ converges if and only if it is Cauchy.*

Since $[-M, M]$ is compact:

Proposition 8.6. *Any bounded sequence in $\mathbb{R}$ has convergent subsequences.*

Proposition 8.7. *A monotonic sequence (e.g., $a_n \leq a_{n+1}$) converges if and only if it is bounded. In this case,*

$$\lim_{n\to\infty} a_n = \sup\{a_n\}.$$

Sometimes we write $a_n \to \pm\infty$; this can be interpreted as convergence in the compactification $\mathbb{R} \cup \{\pm\infty\}$. Such a sequence is still said to diverge.

Definition 8.8. *If (a_n) is bounded, then $M_n = \sup\{a_k : k \geq n\}$ is decreasing. We define the* ***limsup***

$$\limsup a_n := \lim_{n\to\infty} M_n$$

which is the largest limit of a convergent subsequence of (a_n).

We can do the same in the other direction:

Definition 8.9. *If (a_n) is bounded, then $m_n = \inf\{a_k : k \geq n\}$ is increasing. We define the* ***liminf***

$$\liminf a_n := \lim_{n\to\infty} m_n$$

which is the smallest limit of a convergent subsequence of (a_n).

Example 8.10. *The sequences*

$$a_n = \sin\left(\sqrt{n}\pi\right)$$

and

$$a_n = (-1)^n \left(1 + \frac{1}{n}\right)$$

both have $\limsup a_n = 1$ *and* $\liminf a_n = -1$.

Recall:

Proposition 8.11. *A series $\sum a_n$ converges if and only if its partial sums $s_n = \sum_{k=1}^{n} a_k$ form a convergent sequence. In this case, we write $\lim_{n\to\infty} a_n$ for the limit of the sequence.*

Proposition 8.12. *If $\sum a_n$ converges, then $a_n \to 0$. However, the converse is not true: for example, the series $\sum \frac{1}{n}$ diverges, even though $\frac{1}{n} \to 0$.*

Proof. Follows from the Cauchy criterion for the sequence (S_n): $|s_n - s_{n-1}| \to 0$ as $n \to \infty$. □

Proposition 8.13. *For $a_n \geq 0$, the series $\sum a_n$ converges if and only if the partial sums are bounded (since s_n is increasing).*

Proposition 8.14 (Comparison Criterion)**.** *If $0 \leq a_n \leq b_n$ and $\sum b_n$ converges, then $\sum a_n$ converges. Conversely, if $\sum a_n$ diverges, then $\sum b_n$ must also diverge.*

Proposition 8.15. *The geometric series $\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$ converges if and only if $|x| < 1$. It does not converge if $|x| \geq 1$, because the terms do not tend to 0.*

Theorem 8.16. *The series $\sum_{n=1}^{\infty} \frac{1}{n^\alpha}$ converges if and only if $\alpha > 1$.*

Proof. The proof follows from a comparison argument:

$$\frac{2^k}{2^{(k+1)\alpha}} \leq \sum_{n=2^k+1}^{2^{k+1}} \frac{1}{n^\alpha} \leq \frac{2^k}{(2^k)^\alpha},$$

which simplifies to the following inequality for the sum:

$$2^{-\alpha} \sum_{k=0}^{m} 2^{(1-\alpha)k} \leq \sum_{n=2}^{2^{m+1}} \frac{1}{n^\alpha} \leq \sum_{k=0}^{m} 2^{(1-\alpha)k}.$$

This forms a geometric series, and the partial sums are bounded if and only if $2^{1-\alpha} < 1$, which occurs when $\alpha > 1$. □

Proposition 8.17. *The number e is defined as*

$$e := \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = \sum_{k=0}^{\infty} \frac{1}{k!}.$$

Proof. This equality comes from applying the binomial theorem to $\left(1 + \frac{1}{n}\right)^n$, and showing that for fixed k, the binomial coefficient $\binom{n}{k}\left(\frac{1}{n}\right)^k$ increases with n, while $\frac{1}{k!}$ remains constant as $n \to \infty$. □

Proposition 8.18. *e is irrational.*

Proof. Denote the partial sum by $\sum_{k=0}^{n} \frac{1}{k!} = \frac{p_n}{n!}$. Then,

$$e - \frac{p_n}{n!} \in \left(0, \frac{1}{n!}\right),$$

which implies that e cannot be a rational multiple of $\frac{1}{n!}$ for any n, and thus e is irrational. □

Definition 8.19. *A series is said to be **absolutely convergent** if $\sum |a_n|$ converges.*

Proposition 8.20. *If $\sum |a_n|$ converges, then $\sum a_n$ also converges. This can be shown using the Cauchy criterion:*

$$|s_n - s_m| = \left|\sum_{m+1}^{n} a_k\right| \leq \sum_{m+1}^{n} |a_k|.$$

The converse is not true.

Proposition 8.21. *Consider an **alternating series**: if a_n has the same sign as $(-1)^n$, $|a_n|$ is decreasing with n, and $a_n \to 0$, then $\sum a_n$ converges.*

Proof. Proved by showing that the odd and even partial sums increase and decrease towards a common limit. □

Example 8.22.

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - ... \log 2, 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - ... = \arctan(1) = \frac{\pi}{4}.$$

In general, absolutely convergent series can be safely rearranged ($\sum a_{\varphi(n)} = \sum a_n$), multiplied, etc.; others, not always.

Proposition 8.23 (The Root Test)**.** *If $\limsup |a_n|^{\frac{1}{n}} < 1$, then the series $\sum a_n$ converges absolutely (comparison with a geometric series). If $\limsup |a_n|^{\frac{1}{n}} > 1$, then the series $\sum a_n$ diverges (since a_n does not approach 0).*

This test is particularly useful for power series!

Definition 8.24. *The radius of convergence of the power series $\sum a_n x^n$ is given by:*

$$R = \frac{1}{\limsup |a_n|^{\frac{1}{n}}} \in [0, \infty].$$

Theorem 8.25.

- *The power series $\sum a_n x^n$ converges pointwise for all $x \in \mathbb{C}$ such that $|x| < R$.*
- *The series converges uniformly on the closed disk $\overline{B_r(0)} = \{x \in \mathbb{C} : |x| \leq r\}$ for all $r < R$ (but not necessarily on the open disk $B_r(0)$).*
- *Therefore, the function $f(x) = \sum a_n x^n$ is continuous on $B_R(0) = \{x \in \mathbb{C} : |x| < R\}$.*
- *The series diverges for $|x| > R$; at $|x| = R$, the series may either converge or diverge.*

Proof.

- By the root test, we have:

  $$\limsup |a_n x^n|^{\frac{1}{n}} = |x| \limsup |a_n|^{\frac{1}{n}} = \frac{|x|}{R}.$$

  Hence, the series converges for $|x| < R$ and diverges for $|x| > R$.

- Uniform convergence: If $|x| \leq r$, then:

  $$|f(x) - \sum_{k=0}^{n} a_k x^k| = \left| \sum_{k=n+1}^{\infty} a_k x^k \right| \leq \sum_{k=n+1}^{\infty} |a_k| r^k.$$

  The series $\sum |a_n| r^n$ converges by the root test, so $\epsilon_n \to 0$. Therefore, the supremum of the partial sums satisfies:

  $$\sup_{|x| \leq r} \left| f(x) - \sum_{k=0}^{n} a_k x^k \right| \leq \epsilon_n \to 0,$$

  implying uniform convergence.

- Since the partial sums are continuous and the series converges uniformly, the function $f(x)$ is continuous on the closed disk $\{|x| \leq r\}$ for all $r < R$. Thus, $f(x)$ is continuous on $\bigcup_{r<R} \overline{B_r(0)} = B_R(0)$.

□

Example 8.26.

- *The series* $\sum_{n=1}^{\infty}(-1)^{n+1}\frac{x^n}{n}$ *represents the function* $\log(1+x)$ *for* $|x|<1$. *Since* $\lim_{n\to\infty} n^{\frac{1}{n}} = 1$, *the radius of convergence is* $R = 1$. *The series converges at* $x = 1$ *(by the alternating series test) and diverges at* $x = -1$.
- *The series* $\sum_{n=0}^{\infty}\frac{x^n}{n!}$ *represents the exponential function* $\exp(x)$ *and converges everywhere. The radius of convergence is* $R = \infty$. *Indeed, we have* $n! > \left(\frac{n}{2}\right)^{\frac{n}{2}}$, *so* $(n!)^{\frac{1}{n}} > \left(\frac{n}{2}\right)^{\frac{1}{2}} \to \infty$.

Remark 8.27. *Power series form a ring (they can be added and multiplied). Properties of sums and products of numerical series imply that, where the series converge, the sum and product of the series are equal to the sum and product of the corresponding functions.*

8.3 Differentiation in One Variable

Definition 8.28. *A function* $f : [a, b] \to \mathbb{R}$ *is differentiable at* x *if the limit*

$$\lim_{t\to x}\frac{f(t)-f(x)}{t-x} = f'(x)$$

exists. That is, for every $\epsilon > 0$, *there exists a* $\delta > 0$ *such that for all* t *satisfying* $0 < |t - x| < \delta$, *we have*

$$\left|\frac{f(t)-f(x)}{t-x} - f'(x)\right| < \epsilon.$$

Proposition 8.29. *If* f *is differentiable at* x, *then* f *is continuous at* x.

Proof. We have

$$f(t) - f(x) = \frac{f(t)-f(x)}{t-x}\cdot(t-x).$$

As $t \to x$, we know that $\frac{f(t)-f(x)}{t-x} \to f'(x)$, and $(t-x) \to 0$. Since multiplication is continuous, it follows that

$$f(t) - f(x) \to f'(x)\cdot 0 = 0.$$

Hence, $f(t) \to f(x)$ as $t \to x$, proving continuity. □

Remark 8.30. *The converse is false. For example, the function* $f(x) = |x|$ *is continuous at* 0, *but not differentiable there.*

The usual rules of differentiation hold: $(f+g)' = f' + g'$, $fg = f'g + fg'$, and $(f\circ g)'(x) = f'(g(x))g'(x)$.

Example 8.31.

- $f(x) = x \sin\left(\frac{1}{x}\right)$ *for* $x \neq 0$, $f(0) = 0$. *For* $x \neq 0$, *we have*

$$f'(x) = \sin\left(\frac{1}{x}\right) - \frac{1}{x}\cos\left(\frac{1}{x}\right).$$

  This function is continuous but not differentiable at 0, *since* $\lim_{x\to 0} \frac{f(x)}{x}$ *does not exist.*

- $g(x) = x^2 \sin\left(\frac{1}{x}\right)$, *with* $g(0) = 0$, *is differentiable at* 0 *(with* $g'(0) = 0$*), but* g' *is not continuous at* 0.

- $f(x) = \sum_{n=1}^{\infty} \frac{1}{n^2} \sin(n!x)$ *is continuous (since the series converges uniformly, as* $\sum \frac{1}{n^2}$ *converges), but nowhere differentiable.*

Theorem 8.32 (Mean Value Theorem)**.** *If* $f : [a, b] \to \mathbb{R}$ *is differentiable, then there exists a* $c \in (a, b)$ *such that*

$$f(b) - f(a) = f'(c)(b - a).$$

Proof. This follows logically from earlier results:

1. If $f : [a, b] \to \mathbb{R}$ has a local maximum (or minimum) at $x \in (a, b)$, then $f'(x) = 0$ (since $\frac{f(t)-f(x)}{t-x} \geq 0$ for $t \in (x - \delta, x)$, and $\frac{f(t)-f(x)}{t-x} \leq 0$ for $t \in (x, x + \delta)$, implying the limit from both sides is 0).
2. If $f : [a, b] \to \mathbb{R}$ is differentiable and $f(a) = f(b)$, then there exists a $c \in (a, b)$ such that $f'(c) = 0$. This is clear if f is constant; otherwise, apply result (1) to the maximum or minimum of f on $[a, b]$.
3. The Mean Value Theorem can be derived by applying (2) to the function $g(x) = f(x) - \frac{f(b)-f(a)}{b-a}x$.

□

Corollary 8.33 (Mean Value Inequality)**.** *If* $m \leq f'(x) \leq M$ *for all* $x \in (a, b)$, *then*

$$m(b - a) \leq f(b) - f(a) \leq M(b - a).$$

Now, let's generalize:

Theorem 8.34 (Taylor's Theorem)**.** *If* $f : [a, b] \to \mathbb{R}$ *is* n*-times differentiable, then the degree* $(n - 1)$ *Taylor polynomial of* f *at* a *is*

$$P(x) = \sum_{k=0}^{n-1} \frac{f^{(k)}(a)}{k!}(x - a)^k.$$

Moreover, there exists a $c \in (a, b)$ *such that*

$$f(b) = P(b) + \frac{f^{(n)}(c)}{n!}(b - a)^n.$$

Proof. Subtracting $P(x)$ from both sides, we reduce the problem to the case where $f(a) = f'(a) = \cdots = f^{(n-1)}(a) = 0$, and $P(x) = 0$.

Let $g(x) = f(x) - f(b)\frac{(x-a)^n}{(b-a)^n}$, so that $g(a) = g(b) = 0$, and $g'(a) = g'(b) = 0$, and so on up to $g^{(n-1)}(a) = 0$. Applying the Mean Value Theorem to g, we find that there exists a $c_1 \in (a, b)$ such that $g'(c_1) = 0$. Similarly, there exists $c_2 \in (a, c_1)$ such that $g''(c_2) = 0$, and continuing this process, we find a $c_n \in (a, c_{n-1})$ such that $g^{(n)}(c_n) = 0$. This implies that

$$f^{(n)}(c_n) - \frac{n! f(b)}{(b-a)^n} = 0.$$

□

Remark 8.35. *We can compare $f(x)$ to $P(x)$ by applying the theorem to the interval $[a, x]$ instead. As with the Mean Value Inequality, if we have a bound $|f^{(n)}(x)| \leq M$, then we obtain the bound*

$$|f(x) - P(x)| \leq \frac{M(x-a)^n}{n!}$$

for $x \in [a, b]$.

Remark 8.36. *There exist nonzero functions whose Taylor polynomials are all zero! For example, the function*

$$f(x) = \exp\left(-\frac{1}{x^2}\right), \quad f(0) = 0,$$

*is infinitely differentiable ($f \in C^\infty$), and $f^{(k)}(0) = 0$ for all k. The Taylor series of f at 0 converges to 0, but $f(x) \neq 0$ for $x \neq 0$. Most C^∞ functions are not **analytic**, i.e., they cannot be written as power series.*

Let $C^k([a, b], \mathbb{R}) = \{f \mid f^{(k)} \text{ is continuous}\}$, with the norm

$$\|f\|_{C^k} = \sum_{j=0}^{k} \|f^{(j)}\|_\infty.$$

Theorem 8.37. *If $f_n \in C^1$, $f_n \to f$ pointwise, and $f_n' \to g$ uniformly, then $f \in C^1$ and $f' = g$ (and $f_n \to f$ uniformly).*

Proof. Fix $x \neq y \in [a, b]$. By the Mean Value Theorem, we have

$$\frac{f_n(y) - f_n(x)}{y - x} = f_n'(c_n)$$

for some $c_n \in [x, y]$ (or $[y, x]$). The left-hand side converges to $\frac{f(y)-f(x)}{y-x}$ as $n \to \infty$. For the right-hand side, there is a subsequence (c_{n_k}) that converges

to some $c \in [x, y]$. Since f_n' is continuous and $f_n' \to g$ uniformly, we claim that $f_{n_k}'(c_{n_k}) \to g(c)$.

Indeed, fix $\epsilon > 0$. Let $\delta > 0$ be such that $|t - c| < \delta \implies |g(t) - g(c)| < \frac{\epsilon}{2}$. Let N be such that for $n \geq N$, we have $\sup |f_n' - g| < \frac{\epsilon}{2}$, and for $n_k \geq N$, we have $|c_{n_k} - c| < \delta$. Then for $n_k \geq N$,

$$|f_{n_k}'(c_{n_k}) - g(c)| < \epsilon.$$

Thus, taking the limit as $n \to \infty$ in the equation $\frac{f_n(y) - f_n(x)}{y - x} = f_n'(c_n)$, we find that there exists $c \in [x, y]$ such that

$$\frac{f(y) - f(x)}{y - x} = g(c).$$

Taking the limit as $y \to x$, the right-hand side converges to $g(x)$ by the continuity of g, and since $|c - x| \leq |y - x|$, we conclude that f is differentiable at x and $f'(x) = g(x)$. Since g is continuous, it follows that $f \in C^1$.

Finally, the Mean Value Inequality implies that $|f_n(a) - f(a)| + |x - a| \sup |f_n' - f'| \leq b - a$, which gives a uniform bound, so $\sup |f_n - f| \to 0$ uniformly. □

Corollary 8.38. *$C^k([a, b], \mathbb{R})$ is a complete metric space.*

Proof. Using the completeness of C^0 (uniform topology), if (f_n) is Cauchy in C^1, then f_n and f_n' are Cauchy in C^0, so there exist uniform limits $f, g \in C^0$ such that $f \in C^1$ and $f' = g$. Therefore, $f_n \to f$ in C^1, proving the case $k = 1$. The same argument applies for higher derivatives when $k > 1$. □

Corollary 8.39. *If $f(x) = \sum a_n x^n$ is a power series with radius of convergence $R = \infty$, then $f(x)$ is C^∞ over $(-R, R)$ and its derivative is given by*

$$f'(x) = \sum n a_n x^{n-1}.$$

Proof. Both $f(x) = \sum a_n x^n$ and its derivative $g(x) = \sum n a_n x^{n-1}$ have the same radius of convergence. Since the partial sums for both series converge uniformly over compact subsets of $(-R, R)$, we have $f \in C^1$ and $f' = g$. This argument can be repeated for successive derivatives to show $f \in C^\infty$. □

8.4 Riemann Integration

The definite integral of continuous functions is a **linear operator**:

$$\int_a^b (f + g)\, dx = \int_a^b f\, dx + \int_a^b g\, dx, \quad \int_a^b cf\, dx = c \int_a^b f\, dx.$$

Define the map

$$I_a^b : C^0([a,b]) \to \mathbb{R}, \quad f \mapsto I_a^b(f) = \int_a^b f\,dx,$$

for each $a < b \in \mathbb{R}$, which satisfies the following axioms:

1. If $f \geq 0$, then $\int_a^b f\,dx \geq 0$ (i.e., if $f \geq g$, then $\int_a^b f\,dx \geq \int_a^b g\,dx$),
2. If $a < c < b$, then $\int_a^b f\,dx = \int_a^c f\,dx + \int_c^b f\,dx$,
3. $\int_a^b 1\,dx = b - a$.

In fact, such a linear map is unique; the difference between different theories of integration lies in how general a class of functions we allow ourselves to integrate.

The Riemann integral is built starting with **step functions**:

$$s(x) : [a,b] \to \mathbb{R},$$

such that there exist points $a = x_0 < x_1 < \cdots < x_n = b$ where $s(x)$ is constant on each interval (x_{i-1}, x_i), with $s(x) = s_i$. (The values of $s(x)$ at x_i do not matter.) Then, using (2) and (3), we define the integral of $s(x)$ as:

$$I(s) = \int_a^b s(x)\,dx = \sum_{i=1}^n s_i(x_i - x_{i-1}).$$

This definition satisfies the required axioms. Next, if $s \leq f \leq S$ for step functions s and S, then

$$\int_a^b s\,dx \leq \int_a^b f\,dx \leq \int_a^b S\,dx.$$

In particular, if $f : [a,b] \to \mathbb{R}$ is bounded, we can fix $a = x_0 < x_1 < \cdots < x_n = b$, and take $s_i = \inf f([x_{i-1}, x_i])$ and $S_i = \sup f([x_{i-1}, x_i])$, giving the lower and upper Riemann sums of f for the given partition of $[a,b]$. Refining (i.e., subdividing further) the partition provides better bounds on f.

Lower and upper Riemann integrals are defined as:

$$I_-(f) = \sup\left\{\int_a^b s\,dx \mid s \leq f \text{ on } [a,b], s \text{ step function}\right\},$$

$$I_+(f) = \sup\left\{\int_a^b S\,dx \mid S \geq f \text{ on } [a,b], S \text{ step function}\right\}.$$

For all bounded $f : [a,b] \to \mathbb{R}$, we have $I_-(f) \leq I_+(f)$.

Definition 8.40. *f is **Riemann integrable**, $f \in R([a,b])$, if $I_+(f) = I_-(f)$. We set $\int_a^b f\,dx = I_\pm(f)$.*

Theorem 8.41. *Continuous functions are Riemann integrable.*

Proof. The key ingredient is **uniform continuity**: for all $\epsilon > 0$, there exists δ such that for $x, y \in [a,b]$, $|x-y| < \delta \implies |f(x) - f(y)| < \epsilon$. This is proved by applying the Lebesgue number lemma to the open cover $[a,b] \subset \bigcup_{c \in \mathbb{R}} f^{-1}((c, c+\delta))$, where there exists $\delta > 0$ such that for $|x-y| = \operatorname{diam}(\{x,y\}) < \delta$, we have $\{x,y\} \subset f^{-1}((c, c+\epsilon))$.

Thus, given $\epsilon > 0$, take δ as in the uniform continuity definition, and split $a = x_0 < x_1 < \cdots < x_n = b$ such that $x_{i+1} - x_i < \delta$ for all i. Then let $s_i = \min f([x_i, x_{i+1}])$ and $S_i = \max f([x_i, x_{i+1}])$ (attained), which satisfy $S_i - s_i < \epsilon$ for all i, and $s_i \le f \le S_i$ on $[x_i, x_{i+1}]$. Let Δ and S be the step functions taking values Δ_i and S_i on $[x_i, x_{i+1})$, respectively. We have $s \le f \le S$ on $[a,b]$, so $I(\Delta) \le I_-(f)$ and $I(S) \ge I_+(f)$; moreover, $S_i - s_i < \epsilon$ for all i, so $I(S) - I(\Delta) < \epsilon(b-a)$. Hence, we conclude that $I_+(f) - I_-(f) < \epsilon(b-a)$ for all $\epsilon > 0$, which implies $I_+(f) = I_-(f)$ and thus $f \in R([a,b])$.

□

Remark 8.42. *Piecewise continuous functions are also integrable, and some more unusual functions are as well. However, for example, the function*

$$f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q}, \\ 0 & \text{if } x \notin \mathbb{Q} \end{cases}$$

is not Riemann integrable because $I_-(f) = 0$ and $I_+(f) = b - a$. The Lebesgue integral allows more general decompositions into "measurable" subsets (rather than just sub-intervals) and can handle more general functions, including unbounded functions, which are never Riemann integrable. For example, for Riemann integration, $\int_0^x \frac{1}{\sqrt{t}}\,dt = \frac{1}{2}\sqrt{x}$ only makes sense as an "improper integral" (i.e., $\lim_{\epsilon \to 0} \int_\epsilon^x$), whereas the Lebesgue integral can handle this and even worse cases.

In fact, Lebesgue gave a characterization of exactly which functions are Riemann integrable: $f \in R([a,b])$ if and only if f is bounded on $[a,b]$ and the set of points where f is discontinuous has Lebesgue measure 0. This means that for all $\epsilon > 0$, there exists an at most countable collection of open intervals I_i such that $E \subset \bigcup I_i$ and $\sum \text{length}(I_i) < \epsilon$.

It is easy to check (do it!) that $R([a,b])$ is a vector space, and the map $I : R([a,b]) \to \mathbb{R}$ is linear and satisfies the above axioms.

Theorem 8.43 (Fundamental Theorem of Calculus)**.** *If f is continuous on $[a,b]$, then $F(x) = \int_a^x f(t)\,dt$ is differentiable, and $F' = f$.*

Proof. We compute

$$\frac{1}{h}(F(x+h) - F(x)) = \frac{1}{h}\int_x^{x+h} f(t)\,dt \to f(x) \text{ as } h \to 0,$$

using the continuity of f at x to estimate the integral for small h.

□

Theorem 8.44. *$I : C^0([a,b]) \to \mathbb{R}$ is continuous with respect to the uniform topology: if $f_n \to f$ uniformly then $\int_a^b f_n\,dx \to \int_a^b f\,dx$. In fact, $\left|\int f\,dx - \int g\,dx\right| \leq \int |f-g|\,dx \leq (b-a)\sup|f-g|$.*

On the other hand, pointwise convergence isn't enough: Let f_n be the isosceles triangle with base length $\frac{1}{n}$ and height $2n$. Then $f_n \to 0$ pointwise, but $\int_0^1 f_n\,dx = 1$ does not imply $\int_0^1 0\,dx = 0$.

Besides the L^∞ norm, $||f||_\infty = \sup|f|$, we have other norms on the vector space $C^0([a,b],\mathbb{R})$, namely:

$$||f||_1 = \int_a^b |f(x)|\,dx,$$

and for $p \geq 1$,

$$||f||_p = \left(\int_a^b |f(x)|^p\,dx\right)^{\frac{1}{p}}.$$

(Triangle inequality follows from Hölder's inequality, see homework.) These are called L^p norms. Since $||f||_p \leq (b-a)^{\frac{1}{p}}||f||_\infty$, the balls for $||\cdot||_p$ contain balls for $||\cdot||_\infty$, and the topologies defined by these metrics are **coarser** than the uniform topology. Also, $(C^0([a,b]), ||\cdot||_p)$ isn't complete; its completion is the Lebesgue space $L^p([a,b])$ (see Math 114).

Example 8.45. *The function*

$$f_n = \begin{cases} 0 & \text{if } x \leq 0, \\ \frac{1}{n}x & \text{if } 0 < x \leq \frac{1}{n}, \\ 1 & \text{if } x > \frac{1}{n} \end{cases}$$

is Cauchy in the L^1 norm, and in fact converges in L^1 to its pointwise limit

$$f = \begin{cases} 0 & \text{if } x \leq 0, \\ 1 & \text{if } x > 0. \end{cases}$$

We have

$$\int_0^1 |f_n - f|\,dx = \frac{1}{2n} \to 0,$$

but $f \notin C^0$.

The L^1 norm is quite natural, but so is the L^2 norm, which comes from the inner product

$$\langle f, g\rangle_{L^2} = \int_a^b f(x)g(x)\,dx,$$

so that $||f||_{L^2} = \sqrt{\langle f, f\rangle}$. (Cauchy-Schwarz: $\langle f, g\rangle = ||f||_{L^2}||g||_{L^2}$ is a special case of Hölder's inequality.)

We now return to the $||\cdot||_\infty$ (uniform topology) and various results about $C^0([a,b])$.

Closed and bounded subsets of $(C^0([a,b]), ||\cdot||_\infty)$ are not compact. In fact, the closed unit ball of an infinite-dimensional normed vector space is never compact, by Riesz's theorem.

Example 8.46. *The function*

$$f_n = \begin{cases} 0 & \text{if } x \le 0, \\ \frac{1}{n}x & \text{if } 0 < x \le \frac{1}{n}, \\ 1 & \text{if } x > \frac{1}{n} \end{cases}$$

has $||f_n||_\infty$, but there does not exist a uniformly convergent subsequence. Even worse, $f_n = \sin(nx)$ doesn't even have a pointwise convergent subsequence on any interval.

So, what kinds of subsets of $(C^0([a,b]), ||\cdot||_\infty)$ are compact (i.e., sequentially compact)? The Ascoli-Arzelà theorem gives the answer: the family $\{f_n\}$ needs to be uniformly bounded and equicontinuous.

Definition 8.47. *A family of functions $F \subset C^0(K)$, where K is a compact metric space (e.g., $[a,b]$), is **equicontinuous** if for every $\epsilon > 0$, there exists $\delta > 0$ such that for all $f \in F$ and all $x, y \in K$,*

$$d(x,y) < \delta \implies d(f(x), f(y)) < \epsilon.$$

Proposition 8.48. *If $f_n \to f \in C^0(K)$ uniformly, then $\{f_n\}$ is bounded in $||\cdot||_\infty$ (i.e., there exists M such that for all n, $||f_n||_\infty \le M$) and equicontinuous.*

Proof. Given $\epsilon > 0$, there exists N such that for all $n \ge N$, $||f_n - f||_\infty < \frac{\epsilon}{3}$. Since f is uniformly continuous (because K is compact), there exists $\delta > 0$ such that if $d(x,y) < \delta$, then $|f(x) - f(y)| < \frac{\epsilon}{3}$.

Thus, for all $n \ge N$, if $d(x,y) < \delta$, we have

$$|f_n(x) - f_n(y)| \le |f_n(x) - f(x)| + |f(x) - f(y)| + |f_n(y) - f(y)| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon.$$

Since $f_1, \ldots, f_N$ are also uniformly continuous, we can ensure that this holds for $n < N$ by choosing a sufficiently small δ. Hence, the family $\{f_n\}$ is equicontinuous. □

Thus, equicontinuity is necessary for the sequential compactness of subsets of $(C^0(K), ||\cdot||_\infty)$.

Theorem 8.49. *If a sequence $\{f_n\} \subset C^0(K)$ is uniformly bounded and equicontinuous, then it has a uniformly convergent subsequence. Hence, a subset of $(C^0(K), ||\cdot||_\infty)$ is compact if and only if it is closed, bounded, and equicontinuous.*

Proof. Let K be a compact metric space. By the result of the previous proposition, the sequence $\{f_n\}$ is bounded and equicontinuous. We apply the diagonal process to obtain a uniformly convergent subsequence.

Since K is compact, it has a countable dense subset $A = \{x_1, x_2, \dots\} \subset K$ (we can cover K with finitely many $\frac{1}{n}$-balls for all n and take all centers). There exists a subsequence of $\{f_n\}$ that converges pointwise at x_1, another subsequence that converges pointwise at x_2, and so on. By the diagonal process, we obtain a subsequence $\{f_{n_k}\}$ that converges pointwise at all points of A.

We now show that $\{f_{n_k}\}$ is uniformly Cauchy (and hence uniformly convergent) using equicontinuity. Given $\epsilon > 0$, there exists $\delta > 0$ such that for all n_{k_1}, n_{k_2} and all $x, y \in K$ with $|x - y| < \delta$, we have $|f_{n_k}(x) - f_{n_k}(y)| < \frac{\epsilon}{3}$.

Let $A' \subset A$ be a finite subset such that $\bigcup_{x_i \in A'} B_\delta(x_i) \supset K$ (the compactness of K ensures this). There exists N such that for all $n_k, n_l \geq N$, we have $|f_{n_k}(x_i) - f_{n_l}(x_i)| < \frac{\epsilon}{3}$ for all $x_i \in A'$.

Since A' is finite and $\{f_{n_k}\}$ is pointwise Cauchy, we can ensure that for all $x \in K$, there exists some $x_i \in A'$ such that $d(x_i, x) < \delta$. Therefore, for all $n_k, n_l \geq N$, we have

$$|f_{n_k}(x) - f_{n_l}(x)| < \epsilon.$$

Thus, $\{f_{n_k}\}$ is uniformly Cauchy and converges uniformly. □

Example 8.50. *If $(f_n) \in C^1([a,b])$ is a bounded sequence in the C^1-norm (i.e., $\sup|f_n| \leq M$ and $\sup|f_n'| \leq M$), then the sequence is equicontinuous (by the mean value inequality). This implies that there exists a subsequence that converges in C^0. The unit ball for the C^0-norm is not compact in C^0, and the unit ball for the C^1-norm is not compact in C^1, but the C^0-closure of the C^1-unit ball is compact in C^0.*

8.5 Stone-Weierstrass Theorem

Theorem 8.51 (Weierstrass)**.** *Polynomials are dense in $C^0([a,b])$, i.e., for every $f \in C^0([a,b])$, there exists a sequence of polynomials $\{P_n\}$ such that $P_n \to f$ uniformly on $[a,b]$.*

Proof. The proof uses convolution and its ability to approximate smooth functions. □

Definition 8.52. *The convolution of two functions is defined by*

$$(f * g)(x) = \int_{s+t=x} f(s)g(t)\,dt = \int_{-\infty}^{\infty} f(x-t)g(t)\,dt = \int_{-\infty}^{\infty} f(s)g(x-s)\,ds.$$

This is well-defined if, for example, f and g are (piecewise) continuous, and one of them is **compactly supported** (i.e., f or g is zero outside some interval $[-M, M]$). This condition avoids improper integrals.

Principle: The function $f * g$ inherits the best properties of both f and g. Specifically, we have

$$||f * g||_\infty \leq ||f||_{L^1} ||g||_\infty,$$

and thus

$$|(f * g)(x+h) - (f * g)(x)| = \int f(x-t)(g(t+h) - g(t))\,dt \leq ||f||_{L^1} ||g_h - g||_\infty,$$

which we will refer to as the $(\star)$ equation, where $g_h(t) := g(t+h)$.

- If g is continuous, then by uniform continuity (on a compact interval, $|g(t+h) - g(t)| < \epsilon$ for all t when $|h| < \delta$), we have $\lim_{h\to 0} ||g_h - g||_\infty = 0$, implying that $f * g$ is continuous.
- If g is continuously differentiable, i.e., $g \in C^1$, then dividing $(\star)$ by h and applying the Mean Value Theorem, along with the uniform continuity of g', shows that $f * g$ is continuously differentiable and $(f * g)' = f * g'$.
- If g is a polynomial of degree d, then $f * g$ is also a polynomial! This is because $g^{(d+1)} = 0$, so $(f * g)^{(d+1)} = f * g^{(d+1)} = 0$, or more directly: $g(x) = \sum_{k=0}^{d} a_k x^k$ implies

  $$(f*g)(x) = \sum_{k=0}^{d} a_k \int f(t)(x-t)^k\,dt = \sum_{k=0}^{d} \sum_{l=0}^{k} (-1)^l \binom{k}{l} a_k x^{k-l} \int f(t)t^0\,dt,$$

  which is clearly a polynomial in x, since $\int f(t)t^0\,dt$ is constant.

Now we can examine approximate identities.

Definition 8.53. *A sequence of functions $K_n \geq 0$ is called an **approximate identity** if*

$$\int K_n\,dx = 1 \quad \textit{and} \quad \forall \delta > 0, \quad \int_{|x|\geq\delta} K_n\,dx \to 0 \textit{ as } n \to \infty.$$

Theorem 8.54. *If f is compactly supported and continuous, and K_n is an approximate identity, then $f * K_n \to f$ uniformly.*

Proof. We have

$$(f * K_n)(x) - f(x) = \int (f(x-t) - f(x))K_n(t)\,dt = \int_{|t|\le\delta} + \int_{|t|\ge\delta}.$$

We estimate each term as follows:

Given $\epsilon > 0$, by uniform continuity of f on its support, there exists a δ (independent of x) such that for $|t| < \delta$, we have

$$|f(x-t) - f(x)| < \frac{\epsilon}{2}.$$

Therefore,

$$\left|\int_{-\delta}^{\delta} (f(x-t) - f(x))K_n(t)\,dt\right| \le \frac{\epsilon}{2}\int_{-\delta}^{\delta} K_n(t)\,dt \le \frac{\epsilon}{2}.$$

For the second term,

$$\left|\int_{|t|\ge\delta} (f(x-t) - f(x))K_n(t)\,dt\right| \le 2||f||_\infty \int_{|t|\ge\delta} K_n(t)\,dt \to 0 \text{ as } n \to \infty,$$

since

$$2||f||_\infty \int_{|t|\ge\delta} K_n(t)\,dt < \frac{\epsilon}{2} \text{ for sufficiently large } n.$$

This shows that there exists N such that for all $n \ge N$, we have

$$|(f * K_n)(x) - f(x)| < \epsilon \quad \text{for all} \quad x.$$

Thus, $f * K_n \to f$ uniformly. □

Example 8.55. *Let*

$$K_n(x) = c_n(1-x^2)^n \quad \textit{for} \quad |x| \le 1, \quad 0 \ \textit{elsewhere,}$$

where $c_n > 0$ is chosen such that $\int_{-1}^{1} K_n\,dx = 1$.

Claim: K_n is an approximate identity.

Proof:

- *For $|x| < \frac{1}{\sqrt{2n}}$, we have $(1-x^2)^n \geq 1 - nx^2 \geq \frac{1}{2}$, so*

$$\int_{-1}^{1} (1-x^2)^n \, dx \geq \int_{-\frac{1}{\sqrt{2n}}}^{\frac{1}{\sqrt{2n}}} (1-x^2)^n \, dx \geq \frac{1}{\sqrt{2n}},$$

  which implies that $c_n \leq \sqrt{2n}$.

- *For $|x| \geq \delta$, we have $(1-x^2)^n \leq (1-\delta^2)^n$, so*

$$\int_{|x|\geq\delta} K_n \, dx \leq 2c_n(1-\delta^2)^n \leq 2\sqrt{2n}(1-\delta^2)^n \to 0 \quad as \quad n \to \infty.$$

This implies the following theorem:

This leads to the following theorem:

Theorem 8.56 (Weierstrass)**.** *For every $f \in C^0([a,b])$, there exists a sequence of polynomials $\{P_n\}$ such that $P_n \to f$ uniformly.*

Proof. By a linear change of variables, we can assume $[a,b] = [0,1]$. Subtracting a degree 1 polynomial from f, we can assume $f(0) = f(1) = 0$. Then extend f to $\mathbb{R}$ by setting $f(x) = 0$ for $x \notin [0,1]$. Let $K_n(x)$ be as defined above, and let $p_n = f * K_n$. Since K_n is an approximate identity and f is compactly supported and continuous, we have $p_n \to f$ uniformly. Moreover, p_n is a polynomial of degree $2n$ on $[0,1]$ because, given that $f = 0$ outside $[0,1]$, the formula $(f * K_n)(x) = \int f(x-t)K_n(t)\,dt$ for $x \in [0,1]$ does not involve the values of K_n outside $[-1,1]$, and $K_n|_{[-1,1]}$ is a polynomial. □

Stone's Theorem generalizes this to other families of functions:

Definition 8.57. *$\mathcal{A} \subset C^0(K)$ is an **algebra** if for all $f, g \in \mathcal{A}$, we have $f+g \in \mathcal{A}$, $cf \in \mathcal{A}$, and $fg \in \mathcal{A}$. Additionally, $\mathcal{A}$ is said to **separate points** if for all $a \neq b \in K$, there exist $f, g \in \mathcal{A}$ such that $f(a) = 1$, $f(b) = 0$, $g(a) = 0$, and $g(b) = 1$.*

Remark 8.58. *The values 0 and 1 are arbitrary. This is equivalent to saying that the map $\mathcal{A} \to \mathbb{R}^2$ defined by $f \mapsto (f(a), f(b))$ is surjective for all $a \neq b$.*

For complex-valued functions, further assume that $\mathcal{A}$ is conjugation-invariant, i.e., if $f \in \mathcal{A}$, then $\overline{f} \in \mathcal{A}$ (equivalently, $\mathrm{Re}(f) \in \mathcal{A}$ and $\mathrm{Im}(f) \in \mathcal{A}$).

Theorem 8.59 (Stone's Theorem)**.** *Let K be a compact metric space and let $\mathcal{A} \subset C^0(K)$ be an algebra that separates points (and is conjugation-invariant in the complex case). Then $\mathcal{A}$ is dense in $(C^0(K), ||\cdot||_\infty)$.*

Remark 8.60. *Weierstrass' theorem is a special case where $K = [a,b]$ and $\mathcal{A}$ consists of polynomials.*

Proof. The uniform closure of $\mathcal{A}$, denoted $\overline{\mathcal{A}}$, is an algebra (since for $f_n \to f$ and $g_n \to g$, we have $f + g = \lim(f_n + g_n)$ and $fg = \lim(f_n g_n)$). Thus, it is sufficient to show that if $\mathcal{A}$ is closed, then $\mathcal{A} = C^0(K)$.

Given $f \in \mathcal{A}$, we know that $\mathcal{A}$ is an algebra and closed, so any polynomial $P(f)$ such that $P(0) = 0$ must be in $\mathcal{A}$. By Weierstrass' theorem, $|x|$ is a uniform limit of polynomials on $[-M, M]$, so $|f| \in \overline{\mathcal{A}} = \mathcal{A}$.

Therefore, for any $f, g \in \mathcal{A}$, we have $\max(f, g) = \frac{f+g+|f-g|}{2} \in \mathcal{A}$, and similarly for $\min(f, g)$.

Now, given $f \in C^0(K)$ and $\epsilon > 0$, we want to show that there exists $h \in \mathcal{A}$ such that $\sup |h - f| \leq \epsilon$. For any $x \in K$, there exists $y \neq x$ such that $\mathcal{A}$ separates points, so there exist functions $g_y \in \mathcal{A}$ such that $g_y(x) = f(x)$ and $g_y(y) = f(y)$. By covering K with open sets $U_{y_1}, \ldots, U_{y_n}$, we can find $h_x := \max(g_{y_1}, \ldots, g_{y_n}) \in \mathcal{A}$ that satisfies $h_x > f - \epsilon$ and $h_x(x) = f(x)$.

By the same reasoning, we can find $h \in \mathcal{A}$ such that $|h - f| < \epsilon$ everywhere. This completes the proof that $\mathcal{A}$ is dense in $C^0(K)$. □

8.6 Fourier Series

We consider continuous 2π-periodic functions $f : \mathbb{R} \to \mathbb{C}$, or equivalently functions on $S^1 = \mathbb{R}/2\pi\mathbb{Z}$, with the L^2 inner product

$$\langle f, g \rangle = \frac{1}{2\pi} \int_0^{2\pi} \overline{f}(x) g(x)\, dx.$$

The complex exponentials $e_n(x) = e^{inx}$, for $n \in \mathbb{Z}$, satisfy the orthonormality condition

$$\langle e_i, e_j \rangle = \delta_{ij}.$$

Definition 8.61. *The Fourier coefficients of f are given by*

$$c_n(f) = \langle e_n, f \rangle = \frac{1}{2\pi} \int_0^{2\pi} e^{inx} f(x)\, dx.$$

This implies that the Fourier series of f is

$$\sum_{n \in \mathbb{Z}} c_n e_n = \sum_{n=-\infty}^{\infty} c_n(f) e^{inx}.$$

Question (Fourier, Dirichlet, Féjer): Does the Fourier series accurately represent f? In other words, does it converge to f?

Definition 8.62. *The space of **trigonometric polynomials** is the vector space of finite linear combinations of e_n.*

Clearly, this is an algebra, complex conjugate-invariant, and separates points of S^1, which is compact. Hence, by the Stone-Weierstrass theorem, trigonometric polynomials are dense in $(C^0(S^1), \|\cdot\|_\infty)$. Consequently, they are also dense in the L^2-norm, where

$$\|f\|_{L^2} = \left(\frac{1}{2\pi}\int_0^{2\pi} |f(x)|^2\,dx\right)^{1/2} \leq \sup|f|.$$

The n-th Fourier partial sum of f, denoted by $f_n = s_n(f)$, is given by

$$f_n = \sum_{k=-n}^{n} c_k e^{ikx} = \sum_{k=-n}^{n} \langle e_k, f\rangle e_k,$$

which is the orthogonal projection of f onto the subspace $V_n = \operatorname{span}(e_{-n}, \ldots, e_n)$ with respect to the inner product $\langle\cdot,\cdot\rangle$.

Indeed, we have

$$\langle e_j, f_n\rangle = \sum_{k=-n}^{n} c_k\langle e_j, e_k\rangle = c_j = \langle e_j, f\rangle,$$

so that $\langle e_j, f - f_n\rangle = 0$ for all $-n \leq j \leq n$.

Thus, for any $g \in V_n$, we have

$$\|f - f_n\|_{L^2} \leq \|f - g\|_{L^2}.$$

This shows that f_n is the point in V_n closest to f in the L^2-norm. The result follows from the fact that $(f - f_n) \perp V_n$, which implies

$$\|f - g\|^2 = \|f - f_n\|^2 + \|f_n - g\|^2 \geq \|f - f_n\|^2.$$

Theorem 8.63 (Parseval's Theorem)**.** *Let $f \in C^0(S^1)$, and let $c_n = \langle e_n, f\rangle$ be the Fourier coefficients of f. Denote by $f_n = \sum_{k=-n}^{n} c_k e_k$ the partial sums of the Fourier series of f. Then:*

1. *$f_n \to f$ in L^2, i.e.,*

$$\|f_n - f\|_{L^2}^2 = \frac{1}{2\pi}\int_0^{2\pi} |f(x) - f_n(x)|^2\,dx \to 0 \quad \text{as} \quad n \to \infty.$$

2. *$\sum_{n\in\mathbb{Z}} |c_n|^2 = \|f\|_{L^2}^2 = \frac{1}{2\pi}\int_0^{2\pi} |f(x)|^2\,dx$ (in particular, $\sum_{n\in\mathbb{Z}} |c_n|^2$ converges, so $c_n \to 0$ as $|n| \to \infty$).*

Proof.

1. Since trigonometric polynomials are dense in $(C^0(S^1), \|\cdot\|_{L^2})$, for any $\epsilon > 0$, there exists N such that for some $g \in V_N$ with $\|f - g\|_{L^2} < \epsilon$. For $n \geq N$, we have $g \in V_N \subset V_n$, and since f_n is the closest point to f in V_n, we have $\|f - f_n\|_{L^2} \leq \|f - g\|_{L^2} < \epsilon$. This shows that $f_n \to f$ in L^2.

2. Since $f_n \in V_n$ and $f - f_n \in V_n^\perp$, we have

$$\|f\|_{L^2}^2 = \|f_n\|_{L^2}^2 + \|f - f_n\|_{L^2}^2.$$

   Also, $\|f_n\|_{L^2}^2 = \left\|\sum_{k=-n}^n c_k e_k\right\|^2 = \sum_{k=-n}^n |c_k|^2$ by orthonormality. Since $\|f - f_n\|_{L^2}^2 \to 0$ by part (1), we conclude that $\sum_{n\in\mathbb{Z}} |c_n|^2 = \|f\|_{L^2}^2$.

□

Corollary 8.64. *If $f, g \in C^0(S^1)$ have the same Fourier series, then*

$$\frac{1}{2\pi}\int_0^{2\pi} |f(x) - g(x)|^2\, dx = \sum_{n\in\mathbb{Z}} |c_n(f) - c_n(g)|^2 = 0,$$

hence $f = g$.

The fact that $f_n \to f$ in L^2 is the best approximation (in the L^2-norm) of f by trigonometric polynomials. Additionally, since trigonometric polynomials are dense in the $\|\cdot\|_\infty$-norm (i.e., uniformly), one might hope that $f_n \to f$ uniformly or at least pointwise. However, this does not always hold.

Proposition 8.65. *There exists $f \in C^0(S^1)$ such that the Fourier series of f does not converge (e.g., $s_n(f)(0)$ is unbounded).*

However, constructing such an example is quite difficult.

Theorem 8.66 (Dirichlet's Theorem)**.** *If f is C^1, then the Fourier partial sums $s_n(f) \to f$ uniformly.*

The proof uses convolution. For periodic functions, the convolution of f and g is defined by

$$(f * g)(x) = \frac{1}{2\pi}\int_0^{2\pi} f(t)g(x - t)\, dt.$$

Note that

$$c_n e_n(x) = \frac{1}{2\pi}\left(\int f(t)e^{int}\, dt\right) e^{inx} = (f * e_n)(x).$$

Thus, the n-th partial sum of the Fourier series is

$$s_n(f) = \sum_{k=-n}^n c_k e_k = f * \left(\sum_{k=-n}^n e_k\right) = f * D_n,$$

where

$$D_n(x) = \sum_{k=-n}^n e^{ikx} = \frac{\sin\left(\left(n + \frac{1}{2}\right)x\right)}{\sin\left(\frac{x}{2}\right)}$$

is the **Dirichlet kernel**. Dirichlet's proof studies this convolution for $f \in C^1$ to prove uniform convergence. The fact that convergence can sometimes fail makes it remarkable that for all $f \in C^0$, f can still be recovered from the partial sums $s_n(f) = f_n = \sum_{k=-n}^{n} c_k e^{ikx}$.

Theorem 8.67 (Féjer's Theorem)**.** *If $f \in C^0(S^1)$, then*

$$\frac{s_0(f) + \cdots + s_{n-1}(f)}{n}$$

converges uniformly to f.

The reason is that this process amounts to convolution with the Féjer kernel

$$F_n(x) = \frac{D_0 + \cdots + D_{n-1}}{n},$$

which approximates the identity (in the sense described above), unlike the Dirichlet kernel D_n.

8.7 Differentiation in Several Variables

Definition 8.68. *Let $U \subset \mathbb{R}^n$ be open, and let $f : U \to \mathbb{R}^m$. We say that f is **differentiable** at $x \in U$ if there exists a linear map $L : \mathbb{R}^n \to \mathbb{R}^m$ such that*

$$\lim_{v \to 0} \frac{|f(x+v) - f(x) - Lv|}{|v|} = 0.$$

*The **differential** of f at x is then $Df(x) = L \in Hom(\mathbb{R}^n, \mathbb{R}^m)$.*

Remark 8.69. *We can also write the approximation*

$$f(x+v) = f(x) + Lv + o(|v|),$$

where $o(|v|)$ denotes a term that is much smaller than $|v|$, i.e., $\frac{o(|v|)}{|v|} \to 0$ as $v \to 0$.

Conceptually, the input of $Df(x)$ is a **tangent vector** to U at x, and the output, $Df(x)v$, is a tangent vector at $f(x)$.

The natural norm on $\mathrm{Hom}(\mathbb{R}^n, \mathbb{R}^m)$ is the operator norm:

$$\|L\| = \sup_{v \neq 0} \frac{|Lv|}{|v|} = \sup \left\{ \frac{|Lv|}{|v|} \mid v \neq 0 \right\}.$$

We say that $f \in C^1(U, \mathbb{R}^m)$ if f is differentiable at every point of U, and the map $Df : U \to \mathrm{Hom}(\mathbb{R}^n, \mathbb{R}^m)$ is continuous.

As a matrix, the entries of $Df(x)$ are the partial derivatives $\frac{\partial f_j}{\partial x_j}$, which represent the derivatives of f_i with respect to x_j (while keeping the other x_k constant). Then, the differential $Df(x)v$ is given by

$$(Df(x)v)_i = \sum_j \frac{\partial f_i}{\partial x_j} v_j.$$

(Proof: Take $v = e_j$ in the definition of the differential.)

Theorem 8.70. *$f \in C^1(U, \mathbb{R}^m)$ if and only if for all i, j, the partial derivatives $\frac{\partial f_i}{\partial x_j}$ exist and are continuous.*

The implication $\implies$ is clear, but the reverse implication $\impliedby$ is more subtle. The existence of $\frac{\partial f_i}{\partial x_j}$ does not necessarily imply the differentiability or even the continuity of f!

Example 8.71. *Consider the function $f(x, y) = \frac{x^3}{x^2+y^2}$ with $f(0, 0) = 0$. We have*

$$f(x, 0) = x, \quad f(0, y) = 0, \quad \frac{\partial f}{\partial x}(0, 0) = 1, \quad \frac{\partial f}{\partial y}(0, 0) = 0.$$

Thus, if $Df(0)$ exists, it maps $(v_1, v_2) \mapsto v_1$. However, along the path $f(t, t) = \frac{t}{2}$, which is not of the form $t + o(|t|)$.

Proof. We will only prove the $\impliedby$ direction. It is sufficient to consider each component of f, i.e., $f = f_i : U \to \mathbb{R}$, one at a time. Applying the mean value theorem successively, for $x \in U$ and $v \in \mathbb{R}^n$ such that the ball $B_{|v|}(x) \subset U$:

$$f(x_1+v_1, \dots, x_n+v_n) = f(x_1+v_1, \dots, x_{n-1}+v_{n-1}, x_n) + \frac{\partial f}{\partial x_n}(x_1+v_1, \dots, x_{n-1}+v_{n-1}, y_n)v_n$$

$$= f(x_1, \dots, x_n) + \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(x_1 + v_1, \dots, x_{j-1} + v_{j-1}, y_j, x_{j+1}, \dots, x_n) \cdot v_j.$$

Here, the first line uses the mean value theorem for $\frac{\partial f}{\partial x_n}$, and the second line applies the mean value theorem successively to $\frac{\partial f}{\partial x_{n-1}}, \dots, \frac{\partial f}{\partial x_i}$.

All of these points are within distance $|v|$ of x, and since the partial derivatives $\frac{\partial f}{\partial x_j}$ are continuous, for $|v| \to 0$, this expression is well-approximated (within $o(|v|)$) by

$$f(x) + \sum_j \frac{\partial f}{\partial x_j}(x)v_j.$$

Thus, f is differentiable, and $Df(x) = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}\right)$, which depends continuously on x.

□

The usual rules of differentiation hold. In particular:

Theorem 8.72 (The Chain Rule)**.** *If g is differentiable at $x \in \mathbb{R}^n$ and f is differentiable at $g(x) \in \mathbb{R}^m$, then $f \circ g$ is differentiable at x, and*

$$D(f \circ g)(x) = Df(g(x)) \circ Dg(x).$$

Proof. We have

$$g(x+v) = g(x) + Dg(x)v + r(v),$$

where $r(v) = o(|v|)$ (i.e., $\lim_{v\to 0} \frac{|r(v)|}{|v|} = 0$). Thus,

$$(f \circ g)(x+v) = f(g(x)+w) = f(g(x)) + Df(g(x))w + o(|w|).$$

Substituting $w = Dg(x)v$, we get

$$(f \circ g)(x+v) = f(g(x)) + Df(g(x)) \cdot Dg(x)v + o(|v|).$$

This completes the proof. □

Note that the mean value theorem does not always hold. For example, for the function $f : \mathbb{R} \to \mathbb{R}^2$, defined by $f(t) = (\cos t, \sin t)$, we have $f(2n) = f(0)$, but $f(0) + 2\pi f'(t)$ does not hold for all $t \in [0, 2\pi]$. However, we do have the following **mean value inequality**:

Theorem 8.73 (Mean Value Inequality)**.** *If $f : U \to \mathbb{R}^m$ is differentiable at every point of the line segment*

$$[a, b] = \{tb + (1-t)a \mid t \in [0, 1]\},$$

then

$$|f(b) - f(a)| \leq |b - a| \sup_{x \in [a,b]} \|Df(x)\|.$$

Proof. Let u be the unit vector in the direction of $f(b) - f(a)$, and let v be the unit vector in the direction of $b - a$. Define $g(t) = \langle u, f(a + tv)\rangle$. Then

$$g'(t) = \langle u, Df(a + tv)v\rangle,$$

so that

$$|g'(t)| \leq \|Df(a + tv)\|.$$

The result then follows from the single-variable mean value inequality for g on the interval $[0, |b - a|]$. □

Now, let us discuss higher-order derivatives. We say that f is C^2 if $Df : U \to \mathrm{Hom}(\mathbb{R}^n, \mathbb{R}^m) \simeq \mathbb{R}^{n\times m}$ is C^1, and so on. The main important fact about higher partial derivatives is:

Proposition 8.74. *If the second partial derivatives $\frac{\partial^2 f}{\partial x_i \partial x_j}$ exist and are continuous, then*

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}.$$

Proof. It is enough to consider the case of $f(x, y)$. For small h and $k \neq 0$, consider

$$\frac{1}{hk}\left(f(x+h, y+k) - f(x+h, y) - f(x, y+k) + f(x, y)\right).$$

Writing this in terms of $g(x, y) = \frac{f(x,y+k)-f(x,y)}{k}$, we have

$$\frac{1}{h}\left(g(x+h, y) - g(x, y)\right).$$

By the mean value theorem for $\frac{\partial g}{\partial x}$, there exists $h_1 \in (0, h)$ such that this is equal to

$$\frac{\partial g}{\partial x}(x+h, y) = \frac{1}{k}\left(\frac{\partial f}{\partial x}(x+h_1, y+k) - \frac{\partial f}{\partial x}(x+h_1, y)\right).$$

By applying the mean value theorem for $\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)$, there exists $k_1 \in (0, k)$ such that this is equal to

$$\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)(x+h_1, y+k_1).$$

By reversing the order of differentiation, we obtain the same result with $h_2 \in (0, h)$ and $k_2 \in (0, k)$. Since the second derivatives are continuous by assumption, taking limits as $h, k \to 0$ gives the result. □

Hence, the **Hessian** matrix $H = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)$ is symmetric and can be interpreted as a symmetric bilinear form on tangent vectors. If $f \in C^2$, then

$$f(x+v) = f(x) + Df(x) \cdot v + \frac{1}{2}H(x)(v, v) + o(|v|^2).$$

8.8 Inverse Function Theorem

Because of the local approximation

$$f(x+v) = f(x) + Df(x)v + r(v),$$

the behavior of $Df(x)$ governs that of f near x. In particular:

- If $Df(x)$ is injective, then f is injective on a sufficiently small neighborhood of x.
- If $Df(x)$ is surjective, then f maps a neighborhood of x surjectively onto a neighborhood of $f(x)$.

When both conditions hold, f is a local diffeomorphism, by the **Inverse Function Theorem**.

Definition 8.75. *A map $f : U \to V$ between open subsets of $\mathbb{R}^n$ is a **diffeomorphism** if it is a homeomorphism and both f and f^{-1} are C^1.*

Theorem 8.76 (Inverse Function Theorem)**.** *Let $p \in E \subset \mathbb{R}^n$ be open, and let $f : E \to \mathbb{R}^n$ be C^1. Suppose $Df(p) : \mathbb{R}^n \to \mathbb{R}^n$ is an isomorphism (i.e., $\det Df(p) \neq 0$). Then f is a local diffeomorphism at p, i.e., there exists a neighborhood U of p such that f is a diffeomorphism between $U \subset E$ and $f(U) \subset \mathbb{R}^n$.*

The proof uses two main ingredients:

1. The mean value inequality: if $\sup ||Df|| \leq M$, then
$$|f(b) - f(a)| \leq M|b - a|.$$

2. The Contraction Mapping Principle: if X is a complete metric space and $\varphi : X \to X$ is a contraction, i.e.,
$$d(\varphi(x), \varphi(y)) \leq \alpha d(x, y) \text{ for some } \alpha < 1,$$
then φ has a unique fixed point.

For the proof of the contraction mapping principle, suppose $x_0 \in X$ and define the sequence $x_{n+1} = \varphi(x_n)$. Then
$$d(x_n, x_{n+1}) \leq \alpha d(x_{n-1}, x_n),$$
so
$$d(x_n, x_{n+1}) \leq \alpha^n d(x_0, x_1),$$
which implies (x_n) is a Cauchy sequence and thus converges to some $x \in X$. Moreover, since $x_{n+1} = \varphi(x_n)$, we have $\lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} x_n = x$, so $\varphi(x) = x$. Uniqueness: if $\varphi(x) = x$ and $\varphi(y) = y$, then
$$d(\varphi(x), \varphi(y)) = d(x, y) \leq \alpha d(x, y),$$
so $x = y$.

Now, we prove the inverse function theorem.

Proof. After a linear change of variables, we can assume $p = 0$, $f(0) = 0$, and $Df(0) = \mathrm{Id}$. Since $f \in C^1$ and Df is continuous, there exists a ball $B_r(0)$ such that $||Df(x) - I|| < \frac{1}{2}$ for $|x| \leq r$.

Now, given $y_0 \in \mathbb{R}^n$, define the map $\varphi(x) = x + (y_0 - f(x))$. The key observation is that $\varphi(x) = x$ if and only if $f(x) = y_0$. For $|x| \leq r$, we have
$$||D\varphi(x)|| = ||I - Df(x)|| \leq \frac{1}{2}.$$

Assume $|y_0| < \frac{r}{2}$. Since $\varphi(0) = y_0$ and $||D\varphi|| \leq \frac{1}{2}$ for $|x| \leq r$, the mean value inequality gives
$$|\varphi(x_1) - \varphi(x_2)| \leq \frac{1}{2}|x_1 - x_2|,$$

and also

$$|\varphi(x)| \leq |y_0| + \frac{|x|}{2} < r,$$

which we will call the $(\star)$ equation. Therefore, φ is a contracting map from $\overline{B_r(0)}$ to itself, so by the Contraction Mapping Principle, there exists a unique fixed point $x_0 \in B_r(0)$ such that $\varphi(x_0) = x_0$. Hence, for all $y_0 \in B_{\frac{r}{2}}(0)$, there exists a unique $x_0 \in B_r(0)$ such that $f(x_0) = y_0$, which we will call $(\star\star)$.

Now, let $V = B_{\frac{r}{2}}(0)$ and $U = f^{-1}(V) \cap B_r(0)$, where U and V are open sets (since f is continuous). By $(\star\star)$, the map $f|_U : U \to V$ is a bijection. Let $g : V \to U$ be the inverse map.

Claim: g is differentiable and $Dg(y) = Df(x)^{-1}$ where $x = g(y)$ $(y = f(x))$.

Proof: Fix $y_0 \in V$ and $x_0 = g(y_0) \in U$. Let $\varphi(x) = x + (y_0 - f(x))$, as before, with $\varphi(x_0) = x_0$. For small $w \in \mathbb{R}^n$ (so that $|y_0 + w| < \frac{r}{2}$), write $g(y_0 + w) = x_0 + v$, so $f(x_0 + v) = y_0 + w$. Then

$$\varphi(x_0 + v) = (x_0 + v) + (y_0 - (y_0 + w)) = x_0 + v - w,$$

where $\varphi(x_0) = x_0$. Since φ is contracting, we have

$$|\varphi(x_0 + v) - \varphi(x_0)| = |v - w| \leq \frac{1}{2}|v|.$$

Thus, $|w| \geq \frac{1}{2}|v|$ by the triangle inequality, implying $|v| \leq 2|w|$. Given $\epsilon > 0$, there exists δ such that for $|w| < \frac{\delta}{2}$,

$$|(y_0 + w) - y_0 - Df(x_0)v| < \frac{\epsilon}{2}|v| \leq \epsilon|w|.$$

Applying $Df(x_0)^{-1}$, we get

$$|Df(x_0)^{-1}w - v| \leq ||Df(x_0)^{-1}|| \cdot |w - Df(x_0)v| < \epsilon||Df(x_0)^{-1}|| \cdot |w|.$$

Recalling that $v = g(y_0 + w) - g(y_0)$, this yields

$$g(y_0 + w) = g(y_0) + Df(x_0)^{-1}w + o(|w|).$$

Thus, g is differentiable and $Dg = Df(x_0)^{-1}$.

□

Now, let's discuss the implicit function theorem.

Theorem 8.77 (Implicit Function Theorem)**.** *Let $E \subset \mathbb{R}^n \times \mathbb{R}^m$ be open, and let $f : E \to \mathbb{R}^m$ be differentiable, where $(x, y) \mapsto f(x, y)$. Write the derivative of f at (x, y) as $Df(x, y) : \mathbb{R}^n \oplus \mathbb{R}^m \to \mathbb{R}^m$ and decompose it as $Df(x, y) = (Df_x, Df_y)$, where $Df_x : \mathbb{R}^n \to \mathbb{R}^m$ corresponds to the first n variables, and $Df_y : \mathbb{R}^m \to \mathbb{R}^m$ corresponds to the last m variables.*

Assume that $f(x_0, y_0) = 0$ and that Df_y is invertible, i.e., $\det Df_y \neq 0$, at $(x_0, y_0) \in E$. Then, there exist open neighborhoods $U \subset \mathbb{R}^n$ of x_0 and $V \subset \mathbb{R}^m$ of y_0 such that for all $x \in U$, there exists a unique $y = g(x) \in V$ such that $f(x, y) = 0$. Moreover, the map $g : U \to V$ defined by $f(x, g(x)) = 0$ for all $x \in U$ is differentiable, and the derivative of g is given by

$$Dg = -(Df_y)^{-1} Df_x.$$

This result follows from the Inverse Function Theorem by considering the map $F : \mathbb{R}^{n+m} \supset E \to \mathbb{R}^{n+m}$ defined by $F(x, y) = (x, f(x, y))$. The Jacobian matrix of F at (x_0, y_0) is

$$DF(x_0, y_0) = \begin{pmatrix} I & 0 \\ Df_x & Df_y \end{pmatrix},$$

which is invertible. Therefore, F has an inverse G in a neighborhood of (x_0, y_0). Near (x_0, y_0), we have $f(x, y) = 0 \iff F(x, y) = (x, 0) \iff (x, y) = G(x, 0)$. Thus, we define $g(x)$ as the second component of $G(x, 0)$.

Given a differentiable map $f : \mathbb{R}^{n+m} \to \mathbb{R}^m$ and a point at which Df is surjective, we can always find a subset of coordinates $(x_i)_{i \in I}$, where $I \subset \{1, \ldots, n+m\}$ and $|I| = m$, such that the corresponding part of Df is invertible. This allows us to apply the Implicit Function Theorem to describe the zero set of f by equations of the form

$$(x_i)_{i \in I} = g(x_j), \quad \text{for } j \notin I.$$

In particular, a **hypersurface** $S \subset \mathbb{R}^n$ is a closed subset that is locally the zero set of a differentiable real-valued function f with $Df \neq 0$. Using the Implicit Function Theorem, S can be locally described as the graph of some differentiable function $x_j = g(x_i)$ for $i \neq j$. For example, a differentiable curve in $\mathbb{R}^2$ can be locally described as the graph of a function $x = f(y)$ or $y = f(x)$.

8.9 Iterated and Riemann Integrals in Several Variables

Let f be a continuous function on an n-cell $D = [a_1, b_1] \times \cdots \times [a_n, b_n] \subset \mathbb{R}^n$. Then, we can define the integral of f over D as:

$$\int_D f = \int_D f \, dx_1 \, dx_2 \cdots dx_n = \int_D f \, |dx|.$$

Why the notation dx? This will become clearer once we discuss differential forms.

There are two ways to express this integral:

- As an iterated integral:

$$\int_{a_1}^{b_1} \left(\int_{a_2}^{b_2} \cdots \left(\int_{a_n}^{b_n} f(x_1, \ldots, x_n) \, dx_n \right) \cdots dx_2 \right) dx_1,$$

which can be written in any order.

- As a Riemann integral: We divide D into small cubes Q_i and bound f between two piecewise constant functions: $\Delta = \Delta_i = \min f(Q_i)$ on $\operatorname{int}(Q_i)$, and $S = S_i = \max f(Q_i)$ on $\operatorname{int}(Q_i)$. We can estimate the integral by:
$$\sum \Delta_i \operatorname{vol}(Q_i) \leq \int_D f|dx| \leq \sum S_i \operatorname{vol}(Q_i).$$
If f is continuous (and hence uniformly continuous), then $\sup |S - \Delta| \to 0$ as $\operatorname{diam}(Q_i) \to 0$, which defines the integral uniquely.

Theorem 8.78 (Fubini's Theorem)**.** *For a continuous function f, the iterated integrals for different orders of integration are all equal.*

If f is only piecewise continuous, integrability still holds if the regions of D where f is continuous are sufficiently regular, e.g., delimited by smooth hypersurfaces. Specifically, when decomposing D into small cubes Q_i, we require that:
$$\sum_{f|_{Q_i} \text{ not } C^0} \operatorname{vol}(Q_i) \to 0 \quad \text{as one subdivides further.}$$

In such cases, $(S_i - \Delta_i)$ does not tend to 0 as the step size tends to 0, but if $\operatorname{vol}(Q_i) \to 0$, we still have:
$$\int_D (S - \Delta)|dx| = \sum (S_i - \Delta_i)\operatorname{vol}(Q_i) \to 0.$$

Thus, we can define integrals over regions of $\mathbb{R}^n$ delimited by hypersurfaces by either:

- Extending f by 0 outside the given region, and integrating the resulting piecewise continuous function.
- Using a change of coordinates (via the implicit function theorem) to make the region of integration an n-cell. This requires a change of variables.

Theorem 8.79. *Let $\varphi : U \to V$ be a diffeomorphism, and let f be continuous on V. Then,*
$$\int_V f(y)|dy| = \int_U f(\varphi(x)) \left|\det D\varphi(x)\right| \, d|x|.$$

We will not prove this here. The geometric intuition is that if Q_i is a small cube containing x, then $\varphi(Q_i)$ is approximately a small parallelepiped containing $\varphi(x)$, with $\operatorname{vol}(\varphi(Q_i)) \sim |\det D\varphi(x)| \cdot \operatorname{vol}(Q_i)$.

We also consider path integrals. Given a path $\gamma \in C^1([0,1], \mathbb{R}^2)$, where $\gamma(t) = (x(t), y(t))$, and a differential 1-form $\omega = p(x,y)\,dx + q(x,y)\,dy$ with p, q continuous, the path integral is defined as:
$$\int_\gamma \omega = \int_\gamma p\,dx + q\,dy = \int_0^1 \left(p(\gamma(t))x'(t) + q(\gamma(t))y'(t)\right) dt.$$

This integral is independent of the parameterization of the path, due to a change of variables and the chain rule. If we reverse the path, i.e., $(-\gamma)(t) = \gamma(1-t)$, then:

$$\int_{-\gamma} \omega = -\int_{\gamma} \omega.$$

For a function $f \in C^1(\mathbb{R}^2, \mathbb{R})$, we define the differential $df = \frac{\partial f}{\partial x}\,dx + \frac{\partial f}{\partial y}\,dy$. Then, the path integral of df is:

$$\int_{\gamma} df = f(\gamma(1)) - f(\gamma(0)).$$

This result generalizes to arbitrary dimensions using the language of differentiable forms.

8.10 Differential Forms

On $\mathbb{R}^n$, the symbols $dx_1, \ldots, dx_n$ can be viewed as the differentials of the coordinate functions $x_1, \ldots, x_n$; they form a basis of $T^* = \mathrm{Hom}(\mathbb{R}^n, \mathbb{R})$, the space of linear forms on the tangent space $T \simeq \mathbb{R}^n$ (i.e., $dx_i(v) = v_i$, the i^{th} component of the vector v). Differential forms are therefore functions that take values in T^*.

We now consider the **exterior powers** $\bigwedge^k T^*$, which is the vector space with a basis $\{\, dx_{i_1} \wedge \cdots \wedge dx_{i_k} \mid i_1 < \cdots < i_k\}$, consisting of elements of the exterior algebra generated by T^*. This is the quotient of the tensor algebra by the relation $dx_i \wedge dx_j = -\,dx_j \wedge dx_i$ (with $\bigwedge^0 = \mathbb{R}$). This implies that $\alpha \wedge \beta = -\beta \wedge \alpha$ and $\alpha \wedge \alpha = 0$ for all 1-forms.

Definition 8.80. *A **k-form** on an open subset $U \subset \mathbb{R}^n$ is a function with values in $\bigwedge^k T^*$:*

$$\omega = \sum_{i_1 < \cdots < i_k} P_{i_1, \ldots, i_k}(x)\, dx_{i_1} \wedge \cdots \wedge dx_{i_k}$$

(also denoted $\sum_{|I|=k} P_I\, dx_I$).

The space of smooth k-forms on $U \subset \mathbb{R}^n$ is denoted by $\Omega^k(U) = C^\infty(U, \bigwedge^k T^*)$. We can multiply k-forms by functions or take exterior products, where:

$$(f\, dx_{i_1} \wedge \cdots \wedge dx_{i_k}) \wedge (g\, dx_{j_1} \wedge \cdots \wedge dx_{j_l}) = (fg)\, dx_{i_1} \wedge \cdots \wedge dx_{i_k} \wedge dx_{j_1} \wedge \cdots \wedge dx_{j_l}$$

This is zero if $I \cap J = \emptyset$ and equals $\pm(fg)\, dx_{I \sqcup J}$ if $I \cap J = \emptyset$.

Definition 8.81. *The **exterior derivative** $d : \Omega^k \to \Omega^{k+1}$ is defined by:*

$$d\left(\sum_I p_I\, dx_I\right) = \sum_{I,j} \frac{\partial p_I}{\partial x_j}\, dx_j \wedge dx_I.$$

Example 8.82.

- *For $k = 0$ to $k = 1$: $df = \sum \frac{\partial f}{\partial x_i}\, dx_i$.*
- *For $\Omega^1(\mathbb{R}^2) \to \Omega^2(\mathbb{R}^2)$: $d(p\, dx + q\, dy) = \left(-\frac{\partial p}{\partial y} + \frac{\partial q}{\partial x}\right) dx \wedge dy$.*

Proposition 8.83. *$d^2 = 0$, i.e., for all $\omega \in \Omega^k$, $d(d\omega) = 0$. This follows from the symmetry of mixed second partial derivatives, $\frac{\partial^2 p_I}{\partial x_j \partial x_k} = \frac{\partial^2 p_I}{\partial x_k \partial x_j}$, and the antisymmetry of the wedge product: $dx_j \wedge dx_k + dx_k \wedge dx_j = 0$.*

We say that ω is **closed** if $d\omega = 0$, and **exact** if $\omega = d\alpha$ for some $\alpha \in \Omega^{k-1}$. The above proposition implies that exact forms are closed.

Theorem 8.84 (Poincaré Lemma)**.** *For a convex open subset $U \subset \mathbb{R}^n$, a k-form $\omega \in \Omega^k$ is exact if and only if ω is closed, for $1 \leq k \leq n$.*

Remark 8.85. *This result leads to the concept of de Rham cohomology, a key invariant in differential topology. The de Rham cohomology groups are defined by:*

$$H^k_{dR}(U) := \frac{Ker(d : \Omega^k(U) \to \Omega^{k+1}(U))}{Im(d : \Omega^{k-1}(U) \to \Omega^k(U))} = \{\text{closed } k\text{-forms}\}/\{\text{exact forms}\}.$$

The Poincaré Lemma implies that $H^k_{dR}(U) = 0$ for a convex $U \subset \mathbb{R}^n$ and $k \geq 1$. However, $H^1_{dR}(\mathbb{R}^2 - \{0\}) \neq 0$, which detects that $\mathbb{R}^2 - \{0\}$ is not simply connected.

Next, let us discuss the pullback of differential forms. If $\varphi : U \to V$ is a smooth map, where $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^m$, then we define the pullback map $\varphi^* : \Omega^k(V) \to \Omega^k(U)$, which satisfies the following properties:

1. For functions ($k = 0$), $\varphi^*(f) = f \circ \varphi$.
2. $\varphi^*(\alpha \wedge \beta) = \varphi^*\alpha \wedge \varphi^*\beta$.
3. $\varphi^*(d\alpha) = d(\varphi^*\alpha)$.

In coordinates, let (x_i) be the coordinates on U and (y_j) the coordinates on V. The pullback of a 1-form is given by:

$$\varphi^*(dy_j) = d(y_j \circ \varphi) = \sum_i \frac{\partial \varphi_j}{\partial x_i}\, dx_i.$$

For a general k-form, the pullback is:

$$\begin{aligned}\varphi^*\left(\sum_J p_J(y)\, dy_{j_1} \wedge \cdots \wedge dy_{j_k}\right) &= \sum_J p_J(\varphi(x))\, d\varphi_{j_1} \wedge \cdots \wedge d\varphi_{j_k} \\ &= \sum_I \det\left(\frac{\partial(\varphi_{j_1}, \ldots, \varphi_{j_k})}{\partial(x_{i_1}, \ldots, x_{i_k})}\right) dx_I.\end{aligned}$$

Specifically, for $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ and $k = n$:

$$\varphi^*(dy_1 \wedge \cdots \wedge dy_n) = (\det D_\varphi)\, dx_1 \wedge \cdots \wedge dx_n.$$

Let us now discuss the integration of differential forms. Given $\omega = \sum_I (x) p_I(x)\, dx_I \in \Omega^k(U)$, we can integrate ω over a k-dimensional submanifold $M \subset U$ parameterized by a smooth map from a k-cell $D \subset \mathbb{R}^k$ to $U \subset \mathbb{R}^n$ (or any other sufficiently nice domain for integration). Specifically, let $\varphi : D \hookrightarrow U$ be such that $M = \varphi(D)$, where $t \mapsto (\varphi_1(t), \ldots, \varphi_n(t))$. Then, the integral is given by:

$$\int_M \omega = \int_D \sum_I p_I(\varphi(t)) \det \left(\left(\frac{\partial \varphi_i}{\partial t_j} \right)_{1 \le j \le k} \right)_{i \in I} d|t|.$$

We can verify that for 1-forms, this expression agrees with the path integral formula:

$$\int_\gamma p_i\, dx_i = \int p_i(\gamma(t)) \frac{dx_i}{dt}\, dt.$$

What this formula means is the following:

- For n-forms on $D \subset U \subset \mathbb{R}^n$, we have:

$$\int_D f\, dx_1 \wedge \cdots \wedge dx_n = \int_D f\, |dx|.$$

- For a general map $\varphi : D^k \to U \subset \mathbb{R}^n$, we have:

$$\int_{\varphi(D)} \omega = \int_D \varphi^* \omega,$$

  where ω is the k-form on $D \subset \mathbb{R}^k$, and the right-hand side gives the usual integral.

We can similarly integrate k-forms over a finite union of parameterized pieces.

Theorem 8.86 (Pullback Formula)**.** *Given a smooth map $\varphi : U \subset \mathbb{R}^m \to V \subset \mathbb{R}^n$, where $\omega \in \Omega^k(V)$ and $M^k \subset U$, we have:*

$$\int_{\varphi(M)} \omega = \int_M \varphi^* \omega.$$

This formula is essentially equivalent to the change of variables formula for the usual integral $\int_D f\, |dx|$, and it implies that the integral $\int_M \omega$ is independent of the way we parameterize M as the image of a map $\varphi : D \to U$ (or a union of pieces), provided all representations are **orientation-preserving** (i.e., we compare $\varphi : D \to U$ with a diffeomorphism $g : D' \to D$, where $D, D' \subset \mathbb{R}^k$, such that $\det(Dg) > 0$ everywhere).

Example 8.87. *Let* $\omega = \frac{x\,dy - y\,dx}{x^2+y^2}$ *on* $\mathbb{R}^2 - \{0\}$*, and let* C_r *be the circle of radius* r *oriented counterclockwise (parametrized by* $(r,0) \to (r,0)$*). Pulling back via* $\varphi : (r,\theta) \mapsto (r\cos\theta, r\sin\theta)$ *(polar coordinates), we have:*

$$\varphi^*\omega = \frac{(r\cos\theta)(r\cos\theta\,d\theta) - (r\sin\theta)(-r\sin\theta\,d\theta)}{r^2} = d\theta.$$

Thus, we obtain:

$$\int_{C_r} \omega = \int_{\{r\}\times[0,2\pi]} \varphi^*\omega = \int_0^{2\pi} d\theta,$$

which is independent of r*.*

Note that $d\omega = 0$ (either by direct calculation or using the fact that $\varphi^*(d\omega) = d(\varphi^*\omega) = d(d\theta) = 0$), meaning ω is closed, but not exact. If there exists a function $f(x,y)$ on $\mathbb{R}^2 - \{0\}$ such that $df = \omega$, then the path integral:

$$\int_{C_r} \omega = \int_{C_r} df = f(r,0) - f(r,0) = 0.$$

However, the path integral is independent of the radius r, which is a manifestation of the fact that $H^1_{dR}(\mathbb{R}^2 - \{0\}) \neq 0$.

This result is a consequence of **Stokes' Theorem**. For a submanifold $M \subset \mathbb{R}^n$ parameterized as $\varphi(D)$, where $D \subset \mathbb{R}^k$ is a k-cell (or another nice domain), we define the boundary $\partial M = (k-1)$-dimensional boundary $\varphi(\partial D)$. For example, if $D = \prod[a_i, b_i]$ is a k-cell, then the boundary consists of $2k$ pieces, each with a suitable orientation.

Theorem 8.88 (Stokes' Theorem)**.** *For all* $w \in \Omega^{k-1}$*, we have:*

$$\int_M d\omega = \int_{\partial M} \omega.$$

Thus, for example, if ω is a closed 1-form on a simply connected domain $U \subset \mathbb{R}^n$, the path integral $\int_\gamma \omega$ is independent of the choice of path γ from a base point x_0 to x.

Indeed, path-independence follows from Stokes' Theorem for the surface S traced by a path homotopy. Specifically, if $d\omega = 0$, then:

$$0 = \int_S d\omega = \int_{\partial S = \gamma - \gamma'} \omega = \int_\gamma \omega - \int_{\gamma'} \omega.$$

Thus, we can define a function $f(x) = \int_\gamma \omega$ for any path $\gamma : x_0 \to x$. By Stokes' Theorem (which is the fundamental theorem of calculus for differential forms), we have:

$$\int_\gamma df = f(x) - f(x_0) = \int_\gamma \omega \quad \forall \text{ path } \gamma.$$

Therefore, we find that $\omega = df$ is exact, which leads to the Poincaré Lemma.

Remark 8.89. *Stokes' Theorem for differential forms in $\mathbb{R}^2$ and $\mathbb{R}^3$ specializes to the classical theorems of multivariable calculus:*

- *$k = 0$: The fundamental theorem of calculus for path integrals.*
- *$k = 1$: Green's Theorem in $\mathbb{R}^2$ and the curl in $\mathbb{R}^3$.*
- *$k = 2$ in $\mathbb{R}^3$: Gauss' (Divergence) Theorem.*

The most useful case for complex analysis is when $D \subset \mathbb{R}^2$ and we have:

$$\int_{\partial D} p\,dx + q\,dy = \int_D \left(\frac{\partial q}{\partial x} - \frac{\partial p}{\partial y}\right) dx \wedge dy.$$

Let's sketch a proof:

Proof. Both sides of Stokes' Theorem obey the pullback formula (using $\varphi^*(d\omega) = d(\varphi^*\omega)$, and $\partial\varphi(M) = \varphi(\partial M)$), so we can change coordinates or perform the pullback by parameterizing M. We can decompose the integral into pieces, either by writing ω as a sum of forms with support contained in subsets that have a single parameterization, or by observing that if $M = M_1 \cup M_2$ with $M_1 \cap M_2 = N \subset \partial M_i$, then ∂M_1 and ∂M_2 contain N with opposite orientations. Thus, we have:

$$\int_M d\omega = \int_{M_1} d\omega + \int_{M_2} d\omega \quad \text{and} \quad \int_{\partial M} \omega = \int_{\partial M_1} \omega + \int_{\partial M_2} \omega.$$

Over a k-cell, we consider each component of $\omega \in \Omega^{k-1}$ separately. For example, for $D = \prod_{i=1}^k [a_i, b_i] = D' \times [a_k, b_k]$, we have:

$$\omega = f\,dx_1 \wedge \cdots \wedge dx_{k-1} \implies d\omega = (-1)^{k-1} \frac{\partial f}{\partial x_k}\, dx_1 \wedge \cdots \wedge dx_{k-1} \wedge dx_k.$$

Thus,

$$\begin{aligned}
\int_D d\omega &= \int_D (-1)^{k-1} \frac{\partial f}{\partial x_k}\, |dx| \\
&= \int_{D'} \left(\int_{a_k}^{b_k} (-1)^{k-1} \frac{\partial f}{\partial x_k}\, dx_1 \ldots dx_{k-1} \right) \\
&= (-1)^{k-1} \int_{D'} \left[f(x_1, \ldots, x_{k-1}, b_k) - f(x_1, \ldots, x_{k-1}, a_k) \right] dx_1 \ldots dx_{k-1}.
\end{aligned}$$

Thus, we find:

$$\int_D d\omega = (-1)^{k-1} \left(\int_{D' \times \{b_k\}} \omega - \int_{D' \times \{a_k\}} \omega \right).$$

Therefore, we conclude that:

$$\int_{\partial D} \omega = \int_D d\omega.$$

□

9 Complex Analysis I

9.1 Complex Differentiability

We study functions $f : U \subset \mathbb{C}, z \mapsto f(z)$ where $U \subset \mathbb{C}$ is open. Writing $z = x + iy$, these are instances of functions from $\mathbb{R}^2$ to $\mathbb{R}^2$, and the notion of continuity is the same. However, we introduce a different (more restrictive) notion of differentiability.

Definition 9.1. *The **complex derivative** of f at $z \in U$ (if it exists) is*

$$f'(z) = \lim_{h \to 0} \frac{f(z+h) - f(z)}{h}$$

that is, $f(z+h) = f(z) + hf'(z) + o(|h|)$.

The key point is that the limit must hold as $h \to 0$ in the complex plane $\mathbb{C}$.

Definition 9.2. *We say that $f : U \to \mathbb{C}$ is **analytic** (or **holomorphic**) if $f'(z)$ exists for all $z \in U$.*

Example 9.3. *Assume f only takes real values, i.e., $f(z) \in \mathbb{R}$ for all $z \in \mathbb{C}$. In this case, the numerator in the definition of the derivative is always real. When h is real, we get $f'(z) \in \mathbb{R}$, while when h is purely imaginary, we get $f'(z) \in i\mathbb{R}$. Therefore, the complex derivative of a function that takes only real values either does not exist or is equal to zero.*

We can treat $f : U \to \mathbb{C}$ as a function of two real variables, $x + iy$. If $f'(z)$ exists, we can consider limits as h is real or purely imaginary. Specifically, we have:

$$f'(z) = \lim_{h \to 0, h \in \mathbb{R}} \frac{f((x+h)+iy) - f(x+iy)}{h} = \frac{\partial f}{\partial x},$$
$$f'(z) = \lim_{ih \to 0, ih \in i\mathbb{R}} \frac{f(x+i(y+h)) - f(x+iy)}{ih} = -i\frac{\partial f}{\partial y}.$$

This leads to the **Cauchy-Riemann equations**:

$$\frac{\partial f}{\partial x} = -\frac{\partial f}{\partial y}.$$

Equivalently, writing $f = u + iv$ for real-valued functions $u = \mathrm{Re}(f)$ and $v = \mathrm{Im}(f)$, the Cauchy-Riemann equations become:

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y},$$
$$\frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}.$$

In other words, the differential $Df(z) : \mathbb{R}^2 \to \mathbb{R}^2$ is of the form

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

This is the matrix representation of complex multiplication by $f'(z) = a + ib$, viewed as an $\mathbb{R}$-linear transformation on $\mathbb{R} \oplus i\mathbb{R} \simeq \mathbb{C}$.

In the language of differentials, the complex-valued 1-form $df = du + idv$ on $U \subset \mathbb{R}^2$ can be written in terms of $dz = dx + idy$ and $d\overline{z} = dx - idy$ as:

$$\begin{aligned} df &= \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy \\ &= \frac{1}{2}\left(\frac{\partial f}{\partial x} - i\frac{\partial f}{\partial y}\right)(dx + idy) + \frac{1}{2}\left(\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y}\right)(dx - idy) \\ &= \frac{\partial f}{\partial z}dz + \frac{\partial f}{\partial \overline{z}}d\overline{z}. \end{aligned}$$

Thus, if $f'(z)$ exists, we have:

$$\begin{aligned} \frac{\partial f}{\partial \overline{z}} &= \frac{1}{2}\left(\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y}\right)(dx + idy) = 0, \\ \frac{\partial f}{\partial \overline{z}} &= \frac{1}{2}\left(\frac{\partial f}{\partial x} - i\frac{\partial f}{\partial y}\right)(dx + idy) = f'(z). \end{aligned}$$

Conversely, if f is real-differentiable at z, then

$$f(z + h) = f(z) + Df(z) + o(|h|) = f(z) + \frac{\partial f}{\partial z}h + \frac{\partial f}{\partial \overline{z}}\overline{h} + o(|h|),$$

so the complex derivative exists if and only if $\frac{\partial f}{\partial \overline{z}} = 0$.

Proposition 9.4. *The following are equivalent:*

$$\begin{aligned} f \text{ is analytic} &\iff f \text{ is differentiable and } Df \in \left\{\begin{pmatrix} a & -b \\ b & a \end{pmatrix} \mid a, b \in \mathbb{R}\right\} = \mathbb{R}^2 \cdot SO(2) \subset M_{2\times 2}(\mathbb{R}), \\ &\iff \frac{\partial f}{\partial \overline{z}} = 0, \\ &\iff \frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y} = 0. \end{aligned}$$

Remark 9.5. *Rescaling and rotating: conformal transformations. Geometrically, conformal transformations of the plane preserve angles between vectors (and orientation). Thus, analytic functions in one variable are conformal mappings (differentiable in two real variables). If you draw a square grid in the plane and map it by f, the resulting curves will meet at right angles everywhere.*

The miracle: Even though analyticity only requires the existence of a complex derivative, it has many far-reaching consequences, which we will explore and prove in the coming classes. Among these:

1. If $f : U \to \mathbb{C}$ is analytic, then it has derivatives of all orders! (In contrast to the real case, where, for example, $f(x) = x^{\frac{7}{3}}$ is only C^2, not C^∞.)
2. The Taylor series expansion of f at any point $z_0 \in U$ is convergent and equal to f over a disk $B_r(z_0) \subset U$; in particular, $f(z_0+h)$ can be expressed as a power series in h (unlike $f(x) = \exp\left(-\frac{1}{x^2}\right)$, which has all derivatives zero at $x = 0$, so its Taylor series converges to 0, not f).
3. Local determination: If $f, g : U \to \mathbb{C}$ are analytic and U is connected, then $f = g$ on U if $f = g$ on any subset of U that has a limit point (e.g., a small ball, or a small real interval, or ...).

And much more! But first, let's look at examples and work out some basic properties.

Polynomials $P(z) \in \mathbb{C}[z]$: A polynomial $P(z) = \sum_{k=0}^n a_k z^k = a_n \prod_{i=1}^n (z - x_i)$ is analytic, and its complex derivative is simply the usual derivative (this follows from the usual rules of differentiation, which also hold in the complex case). By contrast, a polynomial in two variables $P(x, y)$ can be rewritten as a polynomial in z and $\overline{z}$ (set $x = \frac{z+\overline{z}}{2}$, $y = \frac{z-\overline{z}}{2i}$), so $\mathbb{C}[x, y] \simeq \mathbb{C}[z, \overline{z}]$. To check:

$$\frac{\partial}{\partial z}(z^k \overline{z}^l) = k z^{k-1} \overline{z}^l, \quad \frac{\partial}{\partial \overline{z}}(z^k \overline{z}^l) = l z^k \overline{z}^{l-1},$$

so such a polynomial is analytic if and only if it does not contain $\overline{z}$.

9.2 Rational Functions

A rational function $f \in \mathbb{C}(z)$ is of the form

$$f(z) = \frac{P(z)}{Q(z)} = c \frac{\prod(z - \alpha_i)}{\prod(z - \beta_j)},$$

where $P(z)$ and $Q(z)$ are polynomials, and we assume that the zeros α_i and poles β_j are distinct, i.e., $\alpha_i \neq \beta_j$ for all i, j. The zeros of f are located at the α_i's, and the poles are located at the β_j's. The order of a zero or pole corresponds to the multiplicity of the root α_i or β_j in the polynomials P or Q. Rational functions are analytic on their domain, which is the set $\mathbb{C} - \{\text{poles}\}$.

Rational functions can also be viewed as functions on the **Riemann sphere** $S = \mathbb{C} \cup \{\infty\}$, the one-point compactification of $\mathbb{C}$. Specifically, for a rational function $f(z) = \frac{P(z)}{Q(z)}$, there is a unique continuous extension to a map from S to S, where:

- Poles map to ∞,
- ∞ maps to $\lim_{z\to\infty} \frac{P(z)}{Q(z)} \in \mathbb{C} \cup \{\infty\}$.

At $z = \infty$, the function has a pole of order $\deg(Q) - \deg(P)$ if $\deg(Q) > \deg(P)$, and a zero of order $\deg(P) - \deg(Q)$ if $\deg(P) > \deg(Q)$. This implies that as a map $S \to S$, the number of poles equals the number of zeros (counting multiplicities), and is equal to $\max(\deg(P), \deg(Q)) = \deg(f)$.

Note: For any $c \in S$, the equation $f(z) = c$ has exactly $\deg(f)$ solutions, counting multiplicities. This follows because for $c \in \mathbb{C}$, the degree of the polynomial $f(z) - c$ is $\deg(f)$.

Example 9.6.

- *$f(z) = z^2$ has a zero of order 2 at $z = 0$ and a pole of order 2 at $z = \infty$.*
- *$f(z) = \frac{z}{z^2-1}$ has zeros of order 1 at $z = 0$ and $z = \infty$, and poles of order 1 at $z = \pm 1$.*

The fact that rational functions are analytic maps $S \to S$ can be understood near $z = \infty$ by a change of coordinates, $z = \frac{1}{w}$. The function $f(z)$ is analytic near $z = \infty$ if $f\left(\frac{1}{w}\right)$ is analytic near $w = 0$. Similarly, near poles (infinite values), one can consider the function $\frac{1}{f(z)}$.

In more advanced language, S is a Riemann surface, which has an open cover by two subsets:

- $S - \{\infty\} \simeq \mathbb{C}$, and
- $S - \{0\} \simeq \mathbb{C}$.

The change of coordinates $z = \frac{1}{w}$ is analytic, so we can define analytic functions $S \to S$ as functions whose expressions in these coordinates are analytic. However, one does not need this level of abstraction to study rational functions.

Another perspective (hence the term "sphere") is that we can identify S with the unit sphere in $\mathbb{R}^3$ by stereographic projection:

$$S^2 \to \mathbb{C} \cup \infty : (x, y, z) \mapsto \frac{x + iy}{1 - z},$$

for $z < 1$, and $(0, 0, 1) \mapsto \infty$.

Proposition 9.7. *The map described above is a conformal map from $S^2 \to \mathbb{C} \cup \infty$, meaning it preserves angles.*

Thus, rational functions $f(z) = \frac{P(z)}{Q(z)}$ define conformal maps $S^2 \to S^2$, which are analytic functions $S \to S$ of degree $\deg(f)$, and all conformal maps $S \to S$ are given by rational functions. This result will be proved later.

The special case $\deg(f) = 1$ is of particular interest. These are the fractional linear transformations, or Möbius transformations, of the form

$$f(z) = \frac{az + b}{cz + d}, \quad \det \begin{pmatrix} a & b \\ c & d \end{pmatrix} \neq 0.$$

These transformations are homeomorphisms of S to itself, and they form a group under composition. This group corresponds to the automorphisms of the Riemann sphere, and it is isomorphic to $PSL(2,\mathbb{C})$.

Example 9.8.

- $f(z) = \frac{1}{z}$ *maps* $0 \leftrightarrow \infty$*, and* S^1 *is mapped onto itself by* $e^{i\theta} \mapsto e^{-i\theta}$ *(swapping the hemispheres of* S^2*).*
- $f(z) = i\frac{1-z}{1+z}$ *maps the unit disk* $D = \{|z| < 1\}$ *conformally to the upper half-plane* $H = \{Im(z) > 0\}$*, and maps* S^1 *onto* $\mathbb{R} \cup \infty$*. The analytic isomorphism* $D \simeq H$ *is important and useful in various areas of geometry.*

One way to understand the relation between $z \mapsto \frac{az+b}{cz+d}$ and the matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is to note that $\mathbb{CP}^1 = (\mathbb{C}^2 - \{0\})/(z_1, z_2) \sim (\lambda z_1, \lambda z_2)$ for all $\lambda \in \mathbb{C}^*$. This can be mapped to S, the set of one-dimensional subspaces of $\mathbb{C}^2$, by the map $[z_1, z_2] \mapsto \frac{z_1}{z_2}$. The point $z \in \mathbb{C}$ maps to $[z, 1]$, and ∞ maps to $[1, 0]$. The matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ acts by $[z, 1] \mapsto [az + b, cz + d]$.

Since $\lambda \cdot \mathrm{Id}$ acts trivially, we find that the automorphism group of the Riemann sphere is

$$\mathrm{Aut}(S) = PGL(2,\mathbb{C}) \simeq SL(2,\mathbb{C})/\pm I.$$

This group acts simply transitively on triples of distinct points in S. Specifically, for any distinct points $a_1, a_2, a_3 \in S$ and $b_1, b_2, b_3 \in S$, there exists a unique $f \in \mathrm{Aut}(S)$ such that $f(a_i) = b_i$ for all i.

9.3 Power Series

Consider the power series $f(z) = \sum_{n=0}^{\infty} a_n z^n$ (centered at $z = 0$, or more generally, $\sum_{n=0}^{\infty} a_n (z - z_0)^n$ centered at z_0). Recall the radius of convergence is given by

$$\frac{1}{R} = \limsup_{n\to\infty} |a_n|^{1/n},$$

where $R \in [0, \infty]$.

For $|z| < R$, the series converges (absolutely: $\sum |a_n||z|^n$ converges) by the root test. Specifically, we have

$$\limsup_{n\to\infty} |a_n z^n|^{1/n} = \frac{|z|}{R} < 1,$$

which implies that the series behaves similarly to a geometric series. For $|z| > R$, the series diverges, and for $|z| = R$, the behavior depends on the specific series.

The convergence is uniform on any smaller disk $\overline{D}_r = \{|z| \le r\}$ for $r < R$. This can be shown using the Weierstrass M-test:

$$\sup_{z\in\overline{D}_r} |a_n z^n| = |a_n| r^n, \quad \sum |a_n| r^n \text{ converges for } r < R.$$

Thus, $\sum a_n z^n$ converges uniformly on $\overline{D}_r$. By the Cauchy criterion for partial sums $s_n = \sum_{k=0}^{n} a_k z^k$, for $n > m \leq N$, we have

$$\sup_{z \in \overline{D}_r} |s_n(z) - s_m(z)| = \sup_{\overline{D}_r} \left| \sum_{k=m+1}^{n} a_k z^k \right| \leq \sum_{k=N+1}^{\infty} |a_k| r^k.$$

As $N \to \infty$, we conclude that $f(z) = \sum_{n=0}^{\infty} a_n z^n$ is continuous on $\overline{D}_r$.

The series $g(z) = \sum_{n=0}^{\infty} n a_n z^{n-1}$ has the same radius of convergence as f. The partial sums $s_n(z)$ are analytic, and the partial sums converge uniformly to f, with $s_n'(z) \to g(z)$ uniformly on $\overline{D}_r$ for all $r < R$.

Theorem 9.9. *Let $f(z) = \sum_{n=0}^{\infty} a_n z^n$. Then f is analytic on $\overline{D}_r$, and its derivative is given by*

$$f'(z) = g(z) = \sum_{n=0}^{\infty} n a_n z^{n-1}.$$

Proof. We work on the smaller disk $\overline{D}_r$ with $r < R$, where uniform convergence holds. From the theory of real functions, we know that if the partial sums $s_n \to f$ uniformly, then $s_n'(z) \to g(z)$ uniformly, which implies $f'(z) = g(z)$. For power series, there is an easier proof using mean value inequalities. Since $s_n''(z)$ also converges uniformly on $\overline{D}_r$, we have a uniform bound on $|s_n''(z)| < M$ for all n and for all $z \in \overline{D}_r$. Therefore, for z and $z + h \in \overline{D}_r$, mean value inequalities yield

$$|s_n(z + h) - s_n(z) - s_n'(z)h| \leq \frac{1}{2} M |h|^2.$$

Taking the limit as $n \to \infty$, we obtain

$$|f(z + h) - f(z) - g(z)h| \leq \frac{1}{2} M |h|^2,$$

which shows that $f'(z) = g(z)$. □

Example 9.10. *Consider the power series*

$$\sum_{n=0}^{\infty} z^n = \frac{1}{1 - z},$$

which has radius of convergence $R = 1$. For $|z| = 1$, the series diverges since the terms do not tend to zero, but the right-hand side makes sense for $z \neq 1$. In fact, power series expansions can be obtained around any disk that does not contain the pole at $z = 1$. For example:

- *Around $z_0 = -1$, we have*

$$\frac{1}{1-z} = \frac{1}{2-(z+1)} = \sum_{n=0}^{\infty} \frac{(z+1)^n}{2^{n+1}},$$

  so $R = 2$.

- *Around $z_0 = 2$, we have*

$$\frac{1}{1-z} = \frac{-1}{1+(z-2)} = \sum_{n=0}^{\infty} (-1)^{n+1}(z-2)^n,$$

  so $R = 1$.

- *Around ∞, we have*

$$\frac{1}{1-z} = \frac{-1/z}{1-1/z} = -\sum_{n=1}^{\infty} \left(\frac{1}{z}\right)^n.$$

The process of extending a series beyond its radius of convergence is called **analytic continuation**. For example, here it yields a rational function defined on $\mathbb{C} \setminus \{1\}$. This technique works for all rational functions (e.g., using partial fractions or the case of $\frac{1}{(z-a)^k}$).

Example 9.11. *Consider the partition generating function*

$$\sum p(n) z^n,$$

where $p(n)$ denotes the number of partitions of n (the number of ways of writing n as a sum of positive integers). This is given by:

$$f(z) = \sum p(n) z^n = (1 + z + z^2 + \cdots)(1 + z^2 + z^4 + \cdots)(1 + z^3 + z^6 + \cdots) \cdots$$

which simplifies to

$$f(z) = \prod_{k=1}^{\infty} \frac{1}{1-z^k}.$$

This series converges for $|z| < 1$, and since there are poles at all complex roots of unity, the series cannot be extended past the unit circle.

Example 9.12. *The exponential function $\exp(z) = \sum_{n=0}^{\infty} \frac{z^n}{n!}$ has radius of convergence $R = \infty$, so it converges for all $z \in \mathbb{C}$.*

By algebraic manipulation, we have the identity $\exp(z+w) = \exp(z)\exp(w)$. In particular,

$$e^{-z} = \frac{1}{e^z}, \quad e^z \neq 0 \quad \text{for all } z \in \mathbb{C}.$$

For $z = x + iy$, we have $e^z = e^x e^{iy}$, where $|e^z| = e^x$ and $\arg(e^z) = y$.

The functions $\cos(z)$ *and* $\sin(z)$ *can be defined as*

$$\cos(z) = \frac{e^{iz} + e^{-iz}}{2}, \quad \sin(z) = \frac{e^{iz} - e^{-iz}}{2i}.$$

These are given by the usual series expansions:

$$\cos(z) = 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \cdots, \quad \sin(z) = z - \frac{z^3}{3!} + \frac{z^5}{5!} - \cdots.$$

Note that $\cos(iy) = \cosh(y)$ *for purely imaginary arguments.*

Since $\exp'(z) = \exp(z) \neq 0$*, the exponential function is a local diffeomorphism at each point. Globally,* exp *is the universal covering map from* $\mathbb{C}$ *to* $\mathbb{C}^*$*.*

What about the logarithm? For $w \in \mathbb{C}$, we want to define $\log(w) = z$ such that $e^z = w$. Such a z exists, but it is not unique, as we can add integer multiples of $2\pi i$. The real part of $\log(w)$ is well-defined and equal to $\log|w|$ (for the usual logarithm on $\mathbb{R}_+$).

In general, the expression $\log(w) = \log|w| + i\arg(w)$ is not well-defined and continuous on $\mathbb{C}^*$. However, it is well-defined and continuous on simply connected subsets of $\mathbb{C}^*$. Thus, we cannot define $\arg(w)$ continuously around 0.

This situation is consistent with the lifting problem for the diagram:

$$\begin{array}{ccc} & & \mathbb{C} \\ & \overset{\nearrow}{} & \big\downarrow \exp \\ U & \xrightarrow[i]{} & \mathbb{C}^\times \end{array}$$

The same issue arises when defining z^a for $a \notin \mathbb{Z}$: we would like to define $z^a = \exp(a \log z)$, but this only works on suitable domains. For example, $\sqrt{z}$ is multivalued ($\pm\sqrt{z}$), and we cannot define a continuous function on a domain that encloses the origin.

However, there are still power series expressions away from the origin. For example:

$$\log(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \cdots, \quad \sqrt{1+z} = 1 + \frac{z}{2} - \frac{z^2}{8} + \cdots \quad R = 1.$$

9.4 Cauchy's Theorem and Integral Formula

We now consider path integrals of complex 1-forms $\omega = f(z)\,dz$: given a continuous function $f : U \to \mathbb{C}$ and a (piecewise) differentiable path $\gamma : [0,1] \to \mathbb{C}$, the path integral is given by

$$\int_\gamma f(z)\,dz = \int_0^1 f(\gamma(t))\gamma'(t)\,dt.$$

Alternatively, we can choose points $z_i = \gamma(t_i)$ along the path, with $\operatorname{diam}(\gamma([t_i, t_{i+1}])) < \epsilon$, then the integral is

$$\int_\gamma f(z)\,dz = \lim_{\epsilon \to 0} \sum_i f(z_i)(z_{i+1} - z_i).$$

Example 9.13. *For a path γ from a to b,*

$$\int_\gamma z^n\,dz = \int_0^1 \gamma(t)^n \gamma'(t)\,dt = \frac{1}{n}\left(b^{n+1} - a^{n+1}\right).$$

For a power series, $f(z) = \sum a_n z^n$, if γ is entirely contained within the disc of convergence, it follows that

$$\int_\gamma f(z)\,dz = F(b) - F(a),$$

where $F(z) = \sum \frac{a_n}{n+1} z^{n+1}$: indeed, $F'(z) = f(z)$, and the equality follows from the fundamental theorem of calculus.

In general, a 1-form on $\mathbb{R}^2$ does not need to be exact, and their path integrals need not be path-independent. However, things are much simpler in the analytic setting:

Theorem 9.14 (Cauchy's Theorem)**.** *Let $D \subset \mathbb{C}$ be a bounded region with a piecewise smooth boundary, and let $f(z)$ be analytic on an open set U containing $\overline{D}$. Then,*

$$\int_{\partial D} f(z)\,dz = 0.$$

Proof. Assume f' is continuous: the 1-form $\omega = f(z)\,dz$ is C^1, and $d\omega = df \wedge dz = f'(z) dz \wedge dz = 0$. By Stokes' theorem, we have

$$\int_{\partial D} \omega = \int_D d\omega = 0.$$

□

We'll later show that f analytic implies that f' is continuous. In the meantime, we add the continuity of f' to our working assumptions.

This result holds not only for a simply connected region bounded by a simple closed curve, but also for regions D with holes (e.g., around points where f isn't defined).

Example 9.15. *Let f be analytic on $U - \{z_0\}$, and let γ be a path enclosing z_0 but minus a circle of radius r centered at z_0. Then,*

$$\int_\gamma f(z)\,dz = \int_{C_r} f(z)\,dz,$$

where $C_r = S^1(z_0, r)$ is a circle of radius r centered at z_0, by Cauchy's theorem.

Now assume f is analytic on $U - \{z_0\}$, and $\lim_{z\to z_0}(z - z_0)f(z) = 0$, i.e., f is bounded near z_0. Then,

$$\left|\int_{C_r} f(z)\,dz\right| \leq \sup_{z\in C_r} |f(z)|\cdot\text{length}(C_r) = 2\pi r \sup_{z\in C_r} |f(z)| = 2\pi \sup_{z\in C_r} |(z-z_0)f(z)|.$$

Since this quantity tends to zero as $r \to 0$, and the path integral is independent of r, we obtain:

Theorem 9.16 (Improved Cauchy)**.** *Cauchy's theorem remains true under the weaker assumption that f is defined and analytic in $D - \{z_0\}$, with $z_0 \in int(D)$, and $\lim_{z\to z_0}(z - z_0)f(z) = 0$.*

However, we cannot completely eliminate the assumptions about the behavior of f at z_0.

Example 9.17. *Consider*

$$\int_{S^1(z_0,r)} (z - z_0)^n\,dz = \int_0^{2\pi} \left(re^{i\theta}\right) ire^{i\theta}\,d\theta = \begin{cases} 0 & \text{if } n \neq 1, \\ 2\pi i & \text{if } n = -1. \end{cases}$$

Using this, we derive Cauchy's integral formula:

Theorem 9.18 (Cauchy's Integral Formula)**.** *Let $D \subset \mathbb{C}$ be a bounded region with a piecewise smooth boundary γ, and let $f(z)$ be analytic on an open domain containing $\overline{D}$. If $z_0 \in int(D)$, then*

$$f(z_0) = \frac{1}{2\pi i}\int_\gamma \frac{f(z)\,dz}{z - z_0}.$$

Proof. Since $\int_\gamma \frac{dz}{z-z_0} = 2\pi i$, the formula is equivalent to

$$\frac{1}{2\pi i}\int_\gamma \frac{f(z) - f(z_0)}{z - z_0}\,dz = 0.$$

The differentiability of f at z_0 implies that as $z \to z_0$, we have $\frac{f(z)-f(z_0)}{z-z_0} \to f'(z_0)$, and in particular $(z - z_0)\frac{f(z)-f(z_0)}{z-z_0} \to 0$ (and f is analytic for $z \neq z_0$). The result follows from the improved Cauchy theorem.

□

This is a remarkable result: the values of f at every point inside a closed curve can be determined by evaluating path integrals on γ (assuming f is defined and analytic everywhere in the enclosed region). To emphasize the ability to vary the point of evaluation, Cauchy's integral formula is often written as

$$f(z) = \frac{1}{2\pi i}\int_\gamma \frac{f(w)\,dw}{w - z}.$$

Even more generally:

$$\frac{f^{(n)}(z)}{n!} = \frac{1}{2\pi i}\int_\gamma \frac{f(w)\,dw}{(w-z)^{n+1}}, \quad \forall z \in \mathrm{int}(D), \partial D = \gamma,$$

which implies that all derivatives exist!

Remark 9.19. *If f is given by a power series near z_0, i.e., $f(z) = \sum_{k=0}^\infty a_k(z-z_0)^k$ with $a_k = \frac{f^{(k)}(z_0)}{k!}$, then for $\gamma = S^1(z_0, r)$ (a small circle with r less than the radius of convergence), uniform convergence of the series implies*

$$\frac{1}{2\pi i}\int_{S^1(z_0,r)} \frac{f(w)\,dw}{(w-z_0)^{n+1}} = \sum_{k=0}^\infty \frac{a_k}{2\pi i}\int_{S^1(z_0,r)} \frac{(w-z_0)^k}{(w-z_0)^{n+1}}\,dw = a_n = \frac{f^{(n)}(z_0)}{n!}.$$

Thus, Cauchy's formula implies that $\int_\gamma = \int_{S^1(z_0,r)}$.

However, we haven't yet shown that analytic functions are power series! In fact, the proof of this result uses Cauchy's formula, so we need to work further.

Proposition 9.20. *Suppose $\varphi(w)$ is continuous on $\gamma = \partial D$. Then, for all $n \geq 1$, the function $g_n(z) = \int_\gamma \frac{\varphi(w)\,dw}{(w-z)^n}$ is analytic in the interior of D, and its derivative is given by*

$$g_n'(z) = n\int_\gamma \frac{\varphi(w)\,dw}{(w-z)^{n+1}} = n g_{n+1}(z).$$

Proof. We first prove that g_n is continuous on $\mathrm{int}(D)$. Fix $z_0 \in \mathrm{int}(D)$, with $B_{2\delta}(z_0) \subset D$, and let $z \in B_\delta(z_0)$ (so that z and z_0 are at least δ away from all points of γ). We calculate:

$$\begin{aligned}\frac{1}{(w-z)^n} - \frac{1}{(w-z_0)^n} &= \sum_{k=1}^n \frac{1}{(w-z)^{n-k}(w-z_0)^{k-1}}\left(\frac{1}{w-z} - \frac{1}{w-z_0}\right) \\ &= \sum_{k=1}^n \frac{z-z_0}{(w-z)^{n+1-k}(w-z_0)^k}.\end{aligned}$$

Thus,

$$\begin{aligned}g_n(z) - g_n(z_0) &= \int_\gamma \varphi(w)\left(\frac{1}{(w-z)^n} - \frac{1}{(w-z_0)^n}\right)dw \\ &= (z-z_0)\int_\gamma \varphi(w)\left(\sum_{k=1}^n \frac{1}{(w-z)^{n+1-k}(w-z_0)^k}\right)dw.\end{aligned}$$

Since each term in the sum satisfies $|\cdot| \leq \frac{1}{\delta^{n+1}}$, it follows that:

$$|g_n(z) - g_n(z_0)| \leq |z-z_0| \cdot \left(\sup_{w\in\gamma}|\varphi(w)|\right) \cdot \frac{n}{\delta^{n+1}} \cdot \mathrm{length}(\gamma).$$

Taking $z \to z_0$, this inequality shows that g_n is continuous at z_0. Thus, g_n is continuous on $\mathrm{int}(D)$. Moreover,

$$\frac{g_n(z) - g_n(z_0)}{z - z_0} = \sum_{k=1}^{n} \int_\gamma \frac{\varphi(w)}{(w-z)^{n+1-k}(w-z_0)^k}\, dw \qquad (\star).$$

The continuity result, when applied to $\frac{\partial}{\partial w}\left(\frac{1}{(w-z_0)^k}\right)$, shows that the terms on the right-hand side (RHS) are continuous functions of $z \in \mathrm{int}(D)$. Therefore, the RHS of $(\star)$ is continuous, and its limit as $z \to z_0$ equals $n\int_\gamma \frac{\varphi(w)}{(w-z_0)^{n+1}}\, dw = ng_{n+1}(z_0)$. This gives the existence of:

$$g_n'(z_0) = \lim_{z\to z_0} \frac{g_n(z) - g_n(z_0)}{z - z_0} = \lim_{z\to z_0} (\text{RHS of } (\star)) = ng_{n+1}(z_0).$$

This holds for all $z_0 \in \mathrm{int}(D)$, so g_n is analytic as claimed, and $g_n'(z) = ng_{n+1}(z)$. □

Now, if f is analytic in $U \supset \overline{D}$, then by Cauchy's integral formula, we have:

$$2\pi i f(z) = \int_\gamma \frac{f(w)\, dw}{w - z},$$

which is the expression denoted $g_1(z)$ in the proposition, with $\varphi = f|_\gamma$. The proposition then shows that f is infinitely differentiable, all its derivatives are analytic, and:

$$2\pi i f^{(n)}(z) = n! g_{n+1}(z),$$

i.e.,

$$\frac{f^{(n)}(z)}{n!} = \frac{1}{2\pi i}\int_\gamma \frac{f(w)\, dw}{(w-z)^{n+1}}.$$

This also allows us to lift the extra assumption we've made so far in all proofs using Cauchy's theorem, namely, that f' is continuous.

Proposition 9.21. *If f is analytic, then f' is continuous.*

Proof. If f is analytic in a disc $D \ni z_0$, define $F(z) = \int_{z_0}^{z} f(w)\, dw$, where we choose a path consisting of horizontal and vertical line segments. We don't have the full strength of Stokes' theorem (as we do not yet know if f' is continuous), but we claim it holds for rectangles:

$$\int_{\partial R} f(w)\, dw = 0.$$

Given this, our definition of F makes sense and is path-independent. We now claim that F is analytic, and $F' = f$. Indeed, for $F(z+h) - F(z) = \int_\gamma f(w)\, dw$

where γ is the bottom-right corner path, using the continuity of f, as $h \to 0$, we have:

$$\sup_{\gamma \in w} |f(w) - f(z)| \to 0.$$

Thus,

$$F(z+h) - F(z) = hf(z) + o(|h|),$$

which implies that $F'(z) = f(z)$.

So F is analytic with continuous derivative $F' = f$. We can now apply Cauchy's integral formula to F, so F has derivatives to all orders. In particular, $F''(z) = f'(z)$ is continuous. □

Here's a proof of Cauchy's theorem on rectangles, without assuming f' continuous.

Proof. Here is a proof of Cauchy's theorem on rectangles, without assuming f' is continuous.

Assume $R = R_0$ is a rectangle, f is analytic, and $\int_{\partial R} f(z)\,dz \neq 0$. Cut R into 4 equal rectangles. Then,

$$\int_{\partial R} f(z) = \sum \text{ of 4 path integrals.}$$

Thus, there exists $R_1 \subset R_0$ with $\operatorname{diam}(R_1) = \frac{\operatorname{diam}(R_0)}{2}$ such that:

$$\left| \int_{\partial R_1} f(z)\,dz \right| \geq \frac{1}{4}|I|.$$

We can repeat this process, with $R_0 \supset R_1 \supset \cdots$ and $\operatorname{diam}(R_n) = \frac{\operatorname{diam}(R_0)}{2^n}$, and:

$$\left| \int_{\partial R_n} f(z)\,dz \right| \geq \frac{1}{4^n}|I|.$$

Taking the intersection $\bigcap_{n \in \mathbb{N}} R_n = \{z_0\}$ (a decreasing sequence of non-empty closed subsets in a compact space has a non-empty intersection), we now have:

$$f(z) = f(z_0) + f'(z_0)(z - z_0) + r(z),$$

where $r(z) = o(|z - z_0|)$. Thus,

$$\left|\int_{\partial R_n} f(z)\,dz\right| = \left|\int_{\partial R_n} r(z)\,dz\right| \leq \text{length}(\partial R_n)\cdot \sup_{\partial R_n}|r(z)| = o\left(\frac{1}{4^n}\right),$$

which gives a contradiction.

□

Returning to Cauchy's integral formula for derivatives.

Theorem 9.22 (Cauchy's Integral Formula for Derivatives)**.** *Let $f(z)$ be analytic on $U \subset \mathbb{C}$. Then f has derivatives to all orders in U, all derivatives are analytic, and for $z \in int(D) \subset \overline{D} \subset U$,*

$$\frac{f^{(n)}(z)}{n!} = \frac{1}{2\pi i}\int_{\partial D}\frac{f(w)\,dw}{(w-z)^{n+1}}.$$

We now explore consequences of this formula. First, by bounding the integral on the RHS, we obtain:

Theorem 9.23 (Cauchy's Bound)**.** *If f is analytic in $U \supset \overline{B_r(z_0)}$, then*

$$\left|\frac{f^{(n)}(z_0)}{n!}\right| \leq \frac{1}{R^n}\sup_{w\in S^1(z,R)}|f(w)|.$$

By considering $r < R$ and taking $r \to R$, the result still holds under the weaker assumption that f is continuous on $\overline{B_r(z_0)}$ and analytic in $B_R(z_0)$.

Cauchy's bound has important consequences for **entire functions**, i.e., functions that are analytic on all of $\mathbb{C}$.

Corollary 9.24. *If f is analytic on all of $\mathbb{C}$ ("entire function") and bounded, then f is constant.*

Proof. Apply Cauchy's bound with $R \to \infty$ to obtain $f' = 0$. □

Corollary 9.25. *A nonconstant entire function $f : \mathbb{C} \to \mathbb{C}$ has a dense image, i.e., $\overline{f(\mathbb{C})} = \mathbb{C}$.*

Proof. If $c \in f(\overline{\mathbb{C}})$, then there exists $\epsilon > 0$ such that $|f(z) - c| \geq \epsilon$ for all $z \in \mathbb{C}$, and thus $\frac{1}{f(z)-c}$ is a bounded entire function and hence constant. □

There are even more important consequences for Taylor series of analytic functions.

Corollary 9.26. *The power series $\sum_{n=0}^{\infty}\frac{f^{(n)}(z_0)}{n!}(z-z_0)^n$ (the Taylor series of f at z_0) has radius of convergence $\geq R$, if f is analytic in $B_R(z_0)$.*

Proof. Since Cauchy's bound implies that $\left|\frac{f^{(n)}(z_0)}{n!}\right| \leq \frac{c(r)^{1/n}}{r}$ for all $r < R$, we have $\limsup \leq \frac{1}{r} \implies \leq \frac{1}{R}$. □

Theorem 9.27. *If f is analytic in $B_R(z_0)$, then $f(z) = \sum_{n=0}^{\infty} a_n(z - z_0)^n$, where $a_n = \frac{f^{(n)}(z_0)}{n!}$, for $z \in B_R(z_0)$.*

Proof. By a change of variables, assume $z_0 = 0$. We prove the equality over slightly smaller discs $B_r = \{|z| < r\}$ for all $r < R$; the Taylor series converges by the previous corollary. For $z \in B_r$, write:

$$f(z) = \frac{1}{2\pi i} \int_{S^1(r)} \frac{f(w)\, dw}{w - z},$$

and note that $\frac{1}{w-z} = \frac{1}{w(1-z/w)} = \frac{1}{w} \sum_{n=0}^{\infty} \left(\frac{z}{w}\right)^n$. For fixed $z \in B_r$, this series converges uniformly (by the Weierstrass M-test, as $\sum \left(\frac{|z|}{r}\right)^n$ converges since $|z| < r$). Thus,

$$\frac{1}{2\pi i} \int_{S^1(r)} \frac{f(w)}{w - z}\, dw = \sum_{n=0}^{\infty} \frac{1}{2\pi i} \int_{S^1(r)} \frac{f(w) z^n}{w^{n+1}}\, dw = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} z^n.$$

□

Corollary 9.28. *If $f(z) = \sum a_n z^n$ has radius of convergence R, then it has a singularity (where it cannot be analytically continued) on the circle $\{|z| = R\}$.*

Proof. If there exists an analytic function extending f on an open set $U \supset \overline{B_R(0)}$, then there exists $\epsilon > 0$ such that $B_{R+\epsilon}(0) \subset U$, and thus the radius of convergence would be $\geq R + \epsilon$. □

9.5 Zeroes of Analytic Functions

Corollary 9.29. *Let $f : U \to \mathbb{C}$ be analytic on a connected open set U, and let $z_0 \in U$. If $f^{(n)}(z_0) = 0$ for all n, then $f(z) = 0$ on U. Similarly, if $f^{(n)}(z_0) = g^{(n)}(z_0)$ for all n, then $f = g$ on U.*

Proof. Let $V = \{z \in U \mid f^{(n)}(z) = 0 \, \forall n\}$. By the result on Taylor series, if $z \in V$ and $B_r(z) \subset U$, then f equals its Taylor series at z. This implies that for $z \in V$ and $B_r(z) \subset U$, $f(z)$ is identically zero on $B_r(z)$, so $f^{(n)}(z) = 0$ for all n on $B_r(z)$. Hence, V is open.

Now, let $W = \{z \in U \mid \exists n \text{ such that } f^{(n)}(z) \neq 0\}$, which is the union of sets where some derivative is nonzero. Since W is open, we have $U = V \sqcup W$. Since U is connected and $V = \emptyset$, it follows that $V = U$ and thus $f = 0$ on U. □

The key point is that at a point where $f(z_0) = 0$, the function f vanishes to a finite order (unless $f \equiv 0$), unlike real functions where it may vanish to an infinite or fractional order.

Corollary 9.30. *Let $f : U \to \mathbb{C}$ be analytic on a connected open set U, and suppose f is not identically zero. Then the zeros of f are isolated, i.e., the set $f^{-1}(0)$ has no limit points.*

Proof. If $f(z_0) = 0$, then we can write $f(z) = \sum a_n(z - z_0)^n$, where not all a_n are zero. Let $k = \min\{n \mid a_n \neq 0\}$ (the smallest n for which $a_n \neq 0$). Thus, we have $f(z) = (z - z_0)^k g(z)$, where $g(z) = \sum_{n \geq 0} a_{k+n}(z - z_0)^n$ is analytic in $B(z_0, R) \subset U$ and $g(z_0) = a_k \neq 0$. By continuity, there exists an $\epsilon > 0$ such that if $|z - z_0| < \epsilon$, then $g(z) \neq 0$. Therefore, for $0 < |z - z_0| < \epsilon$, we have $f(z) \neq 0$, so z_0 is an isolated zero of f. □

Remark 9.31. *In the real C^∞ world, there are nonzero functions with nonisolated zeros, such as $f(x) = \exp\left(-\frac{1}{x^2}\right) \sin\left(\frac{1}{k}\right)$ for $x \neq 0$, with $f(0) = 0$.*

Corollary 9.32 (Uniqueness of Analytic Continuation)**.** *Let $f, g : U \to \mathbb{C}$ be analytic on a connected open set U. If $f = g$ on a nonempty open subset of U, or on any subset of U that has a limit point, then $f = g$ on all of U.*

Let's take a look at some other consequences of Cauchy formula, for the space of analytic functions with the uniform topology.

Theorem 9.33. *Let $f_n(z)$ be a sequence of analytic functions on a set U. If $f_n \to f$ **locally uniformly** (i.e., for each $z \in U$, there exists $r > 0$ such that $\overline{B_r(z)} \subset U$ and $f_n \to f$ uniformly on $\overline{B_r(z)}$), then f is analytic on U.*

Proof. Given a closed disk $B \subset U$ where $f_n \to f$ uniformly, and $z \in \text{int}(B)$, we have:

$$f(z) = \lim_{n\to\infty} f_n(z) = \lim_{n\to\infty} \frac{1}{2\pi i} \int_{\partial B} \frac{f_n(w)}{w - z}\, dw = \frac{1}{2\pi i} \int_{\partial B} \frac{f(w)}{w - z}\, dw.$$

Thus,

$$\int_{\partial B} \frac{f_n(w) - f(w)}{w - z}\, dw \to 0 \text{ as } n \to \infty.$$

Therefore, f is analytic on U. □

Theorem 9.34. *If f_n is a sequence of analytic functions on U that converges locally uniformly to f, then the sequence of derivatives f_n' converges locally uniformly to f' and this holds for higher derivatives as well.*

Proof. The proof follows from the same reasoning as the previous theorem, applying the Cauchy formula and the uniform convergence of the functions. □

Theorem 9.35. *Any uniformly bounded sequence of analytic functions f_n on U has a subsequence that converges uniformly on compact sets to an analytic function g.*

Proof. Let $K \subset U$ be compact. There exists $r > 0$ such that the distance from K to the boundary of U is greater than r. For each $z \in K$,

$$|f_n'(z)| = \left| \frac{1}{2\pi i} \int_{S^1(z,r)} \frac{f_n(w)}{(w-z)^2}\, dw \right| \leq \frac{1}{2\pi r} \sup_n |f_n|.$$

Since (f_n) is uniformly bounded, this gives a uniform bound on $|f_n'|$ on K. Thus, (f_n) is uniformly equicontinuous on K, and by the Ascoli-Arzelà theorem, there exists a subsequence that converges uniformly on K. We can ensure uniform convergence on all compact sets by considering a sequence of compact sets K_n such that $\bigcup_n K_n = U$, and using a diagonalization argument to obtain a subsequence that converges uniformly on all of them. □

In real analysis, a standard example of a sequence of continuous functions that is not equicontinuous over $[-a, a]$ for all $a > 0$ is given by

$$f_n(x) = \frac{1}{1 + n^2 x^2}.$$

This sequence has no uniformly convergent subsequence because its pointwise limit is not continuous. These functions can be extended to analytic functions $f_n(z) = \frac{1}{1+n^2 z^2}$, but the theorem above does not apply near 0 because f_n has poles at $z = \pm\frac{1}{n}$, and thus the sequence is not uniformly bounded on any fixed neighborhood of 0.

Besides the powerful results (such as derivatives to all orders, Cauchy's formula, and convergence of Taylor series), there are also more fundamental concepts from real analysis that carry over to the complex case, such as antiderivatives and inverse functions. However, these results come with certain caveats.

Theorem 9.36. *If $f(z)$ is analytic on a simply connected open set $U \subset \mathbb{C}$, then there exists an analytic function $F : U \to \mathbb{C}$ such that $F'(z) = f(z)$.*

Proof. This is because we can define $F(z) = \int_{z_0}^{z} f(z)\, dz$. Cauchy's theorem implies that the choice of path does not matter: given any piecewise differentiable closed loop γ in U, we have

$$\int_\gamma f(z)\, dz = 0.$$

In fact, over discs $B_r(z_0) \subset U$, we can define F by term-by-term integration of the power series expression for f. □

Note that the condition of being simply connected is necessary. For example, the function $f(z) = \frac{1}{z}$ on $C^* = \mathbb{C} \setminus \{0\}$ can only be integrated to $F(z) = \log z$ over a simply connected subset (i.e., a domain that does not allow paths enclosing 0).

Theorem 9.37. *If f is analytic near a, with $f(a) = b$ and $f'(a) = 0$, then there exists an analytic inverse function g defined on a neighborhood of b, such that $g(b) = a$ and $g'(b) = \frac{1}{f'(a)}$.*

Proof. This result is a direct consequence of the inverse function theorem for $f : \mathbb{R}^2 \to \mathbb{R}^2$, along with the observation that if $f'(a) \neq 0$, then the derivative $Df(a)$ is invertible, and its inverse is also complex linear. □

Remark 9.38. *For real functions of one real variable, this result holds on any connected interval where $f'(x) \neq 0$ (implying that f is injective). However, in the complex world, this is not necessarily true, even on simply connected domains. For instance, the logarithm function $\log(z)$ is the inverse function of the exponential function $\exp(z)$, but it is only defined on suitable domains that avoid enclosing 0. Similarly, $\sqrt[n]{z}$ is the inverse function of z^n, but it is defined only on specific domains.*

The inverse function theorem gives us that for $\exp'(z) = e^z$, we have $\log'(z) = \frac{1}{z}$, and the derivative of $z^{1/n}$ is $\frac{1}{n} z^{-(n-1)/n}$. The corresponding power series expansions are:

$$\log(1+z) = \int \frac{dz}{1+z} = z - \frac{z^2}{2} + \frac{z^3}{3} - \dots$$

with radius of convergence $R = 1$, and

$$(1+z)^\alpha = 1 + \alpha z + \frac{\alpha(\alpha-1)}{2} z^2 + \frac{\alpha(\alpha-1)(\alpha-2)}{3!} z^3 + \dots$$

These functions exhibit singularities at $z = 0$—referred to as "branch singularities"—not poles. We will soon study the behavior of analytic functions at isolated singularities, that is, when f is defined on $U \setminus \{z_0\}$, where z_0 is an isolated point of U. However, this analysis will not cover functions like $\log(z)$ or z^α, which are not analytic on the entire domain $D^*(r) = D(r) \setminus \{0\}$.

We will now study the behavior of analytic functions at isolated singularities, where f is defined on $U \setminus \{z_0\}$, with $z_0 \in \text{int}(U)$. However, this analysis will not handle functions such as $\log(z)$ or z^α, which are not analytic on the entire domain $D^*(r) = D(r) \setminus \{0\}$.

9.6 Laurent Series

Laurent series are power series with both positive and negative exponents:

$$f(z) = \sum_{n=-\infty}^{\infty} a_n z^n.$$

The convergence of a Laurent series is best understood by splitting it into two parts: the sum of the non-negative powers of z and the sum of the negative powers of z:

$$\sum_{n\geq 0} a_n z^n$$

which converges when $|z| < R_2$, where

$$R_2 = \frac{1}{\limsup_{n\to\infty} |a_n|^{1/n}}.$$

and

$$\sum_{n<0} a_n z^n$$

which converges when $|z| > R_1$, where

$$R_1 = \limsup_{n\to-\infty} |a_n|^{1/n}.$$

This implies that the Laurent series converges in an annulus, specifically for $R_1 < |z| < R_2$.

It is important to note that the general formal Laurent series does not form a ring. The issue arises because the coefficient of z^n in the product $\left(\sum a_k z^k\right)\left(\sum b_k z^k\right)$ should be $\sum_{k\in\mathbb{Z}} a_k b_{n-k}$, which may not result in a convergent series. However, things work well when the annuli of convergence intersect non-trivially. A better-behaved class of Laurent series consists of those with only finitely many negative powers of z, i.e., $\sum_{n=-N}^{\infty} a_n z^n$ (which is essentially $\frac{1}{z^N}\cdot$ (a power series)). These series actually form a field, known as the field of fractions of the ring of power series.

Theorem 9.39. *If $f(z)$ is analytic in the annulus $A_{R_1,R_2} = \{R_1 < |z| < R_2\}$, then $f(z)$ can be expressed as a Laurent series:*

$$f(z) = \sum_{n=-\infty}^{\infty} a_n z^n,$$

which converges in the annulus A_{R_1,R_2}.

Proof. We will prove this result for slightly smaller annuli $\{r_1 \leq |z| \leq r_2\}$, where $R_1 < r_1 < r_2 < R_2$. Then, using Cauchy's formula for the annulus A_{r_1,r_2} and its boundary $S^1(r_2) - S^1(r_1)$, we have:

$$f(z) = \frac{1}{2\pi i}\int_{S^1(r_2)} \frac{f(w)\,dw}{w-z} - \frac{1}{2\pi i}\int_{S^1(r_1)} \frac{f(w)\,dw}{w-z},$$

for $r_1 < |z| < r_2$. On $S^1(r_2)$ with $|z/w| < 1$, we have:

$$\frac{1}{w-z} = \frac{w^{-1}}{1-z/w} = \sum_{n=0}^{\infty} \frac{z^n}{w^{n+1}},$$

which converges uniformly. On $S^1(r_1)$ with $|w/z| < 1$, we have:

$$\frac{1}{z-w} = \frac{z^{-1}}{1-w/z} = \sum_{n=0}^{\infty} \frac{w^n}{z^{n+1}} = \sum_{n\leq -1} \frac{z^n}{w^{n+1}},$$

which also converges uniformly. Uniform convergence allows us to interchange the sum and the integral, yielding:

$$f(z) = \sum_{n\geq 0} \frac{1}{2\pi i} z^n \int_{S^1(r_2)} \frac{f(w)\,dw}{w^{n+1}} + \sum_{n\leq -1} \frac{1}{2\pi i} z^n \int_{S^1(r_1)} \frac{f(w)\,dw}{w^{n+1}}.$$

This simplifies to:

$$f(z) = \sum_{n\in\mathbb{Z}} a_n z^n,$$

where $a_n = \frac{1}{2\pi i} \int_{S^1(r)} \frac{f(w)\,dw}{w^{n+1}}$, for any $r \in (R_1, R_2)$, since this is independent of r by Cauchy's theorem. □

Corollary 9.40. *Any analytic function on the annulus $\{R_1 < |z| < R_2\}$ can be written as the sum of an analytic function on $\{|z| < R_2\}$ and an analytic function on $\{|z| > R_1\}$.*

9.7 Singularities and Removability

Let f be analytic on $D^*(R) = D(R) \setminus \{0\}$, and express it as a Laurent series:

$$f(z) = \sum_{n\in\mathbb{Z}} a_n z^n.$$

Let $N = \inf\{n \in \mathbb{Z} \mid a_n \neq 0\}$, provided this infimum exists.

1. If $N \geq 0$ (i.e., $a_n = 0$ for all $n < 0$), then f is a power series, and the singularity at 0 is **removable**. In other words, f can be extended to an analytic function on $D(R)$, including at 0.

- If $N = \infty$ (i.e., $a_n = 0$ for all n), then $f \equiv 0$.
- If $N > 0$, then $f(z) = z^N (a_N + \dots)$, so f has an isolated zero of order N at 0.
- If $N = 0$, then $f(0) = a_0 \neq 0$.

1. If $N < 0$ is finite (i.e., there are finitely many negative powers of z in the series), then
$$f(z) = \frac{1}{z^{|N|}}(a_N + \dots) = \frac{g(z)}{z^{|N|}},$$
where $g(z)$ is analytic with $g(0) = a_N \neq 0$. We say that f has a **pole** of order $|N|$ at 0.

2. If $N = -\infty$ (i.e., the negative part of the series has infinitely many terms), then f has an **essential singularity** at 0 (a non-removable singularity other than a pole). For example, $\exp(1/z) = \sum_{n=0}^{\infty} \frac{1}{n!} z^{-n}$ has an essential singularity at 0.

The qualitative differences between the three cases can also be understood without using Laurent series.

Theorem 9.41. *Let f be analytic on $D^*(R)$:*

1. *The singularity at 0 is removable if and only if $f(z)$ is bounded on a neighborhood of 0.*
2. *f has a pole at 0 if and only if $|f(z)| \to \infty$ as $z \to 0$.*
3. *f has an essential singularity if and only if for every $\epsilon > 0$, $f(D^*(\epsilon))$ is dense in $\mathbb{C}$ (equivalently: for every $y \in \mathbb{C} \cup \{\infty\}$, there exists a sequence $z_n \to 0$ such that $f(z_n) \to y$).*

Proof. We will prove these results without using Laurent series.

1. Assume that f is bounded on $\overline{D^*(r)}$ for some $r > 0$. Since f is continuous on $S^1(r)$, we know that the function
$$g(z) = \frac{1}{2\pi i} \int_{S^1(r)} \frac{f(w)\,dw}{w - z}$$
is analytic in $D(r)$. By Cauchy's formula, if $0 < \epsilon < \frac{|z|}{2}$, then:
$$\begin{aligned} \frac{1}{2\pi i} \int_{\partial D} \frac{f(w)\,dw}{w - z} &= \frac{1}{2\pi i} \left(\int_{S^1(r)} - \int_{S^1(z,\epsilon)} - \int_{S^1(0,\epsilon)} \right) \\ &= g(z) - f(z) - \frac{1}{2\pi i} \int_{S^1(0,\epsilon)} \frac{f(w)\,dw}{w - z}. \end{aligned}$$
The last integral tends to 0 as $\epsilon \to 0$ because the integrand is bounded and the length of $S^1(\epsilon)$ tends to 0. Therefore, $g(z)$ is analytic in $D(r)$, and $g(z) = f(z)$ for all $z \in D(r) \setminus \{0\}$. This shows that the singularity at 0 is removable. Conversely, if the singularity is removable, f must be bounded near 0.

2. Assume that $|f(z)| \to \infty$ as $z \to 0$. Let $h(z) = \frac{1}{f(z)}$, which is analytic and bounded in a neighborhood of 0, and hence has a removable singularity

at 0. Thus, $h(z)$ can be extended analytically over 0. Since $|h(z)| \to 0$ as $z \to 0$, h has an isolated zero at $z = 0$, which vanishes to finite order. Therefore, there exists $n \geq 1$ and an analytic function $k(z)$ with $k(0) \neq 0$ such that $h(z) = z^n k(z)$. Consequently, $f(z) = \frac{1}{h(z)} = \frac{g(z)}{z^n}$, where $g(z) = \frac{1}{k(z)}$ is analytic in a neighborhood of 0. Thus, f has a pole of order n. Conversely, if $f(z) = \frac{g(z)}{z^n}$ for some $n \geq 1$, where g is analytic and $g(0) \neq 0$, then there exists a constant $c > 0$ such that $|g(z)| \geq c > 0$ in a neighborhood of 0, and $|f(z)| \geq \frac{c}{|z|^n} \to \infty$ as $z \to 0$. Therefore, f has a pole of order n.

3. If $f(D^*(\epsilon))$ is not dense in $\mathbb{C}$, then there exists a constant c such that $h(z) = \frac{1}{f(z)-c}$ is bounded near 0, and hence h has a removable singularity. Let the extension of h over 0 be denoted again by h. If $h(0) = 0$, then, as in the previous case, h has a zero of finite order $n \geq 1$, and $\frac{1}{h(z)}$ has a pole of order n. Therefore, $f(z) = c + \frac{1}{h(z)}$ has a pole of order n. If $h(0) \neq 0$, then $f(z) = c + \frac{1}{h(z)}$ extends analytically over 0, and the singularity is removable. Thus, if f has an essential singularity, then $f(D^*(\epsilon))$ is dense in $\mathbb{C}$ for all $\epsilon > 0$. Conversely, if $f(D^*(\epsilon))$ is dense in $\mathbb{C}$, then f cannot be bounded and cannot have a pole, so it must have an essential singularity.

□

9.8 Meromorphic Functions

Definition 9.42. *A function f is said to be **meromorphic** in a domain U if it is analytic on $U \setminus \{p_1, \ldots, p_n\}$ and has poles at the points $p_1, \ldots, p_n$ (i.e., no essential singularities).*

If $f : U \setminus \{p_i\} \to \mathbb{C}$ is meromorphic with poles at p_i, then $|f(z)| \to \infty$ as $z \to p_i$. Thus, $\frac{1}{f}$ has a removable singularity at each p_i, where it has a zero of order equal to the pole order of f. Hence, f extends to a function $\hat{f} : U \to S = \mathbb{C} \cup \{\infty\}$ on the Riemann sphere by setting $\hat{f}(p_i) = \infty$. The function $\hat{f}$ is continuous and analytic in the sense that:

- Away from the poles $\{p_i\} = \hat{f}^{-1}(\infty)$, $\hat{f}$ takes values in $\mathbb{C}$ and is analytic.
- Away from the zeros of $\hat{f}$, $\frac{1}{\hat{f}(z)}$ is analytic (this is the analytic extension of $\frac{1}{f}$ over the removable singularity at p_i).

From these considerations, we deduce the following facts:

- The zeros and poles of a non-identically zero meromorphic function are isolated.
- If f and g are analytic on U, and g is not identically zero, then $\frac{f}{g}$ is meromorphic on U. If f and g have no common zeros, the zeros of $\frac{f}{g}$ coincide with the zeros of f, and the poles of $\frac{f}{g}$ coincide with the zeros of

g. If there is a common zero, the highest order zero (or pole) dominates, and we factor out powers of $(z - z_0)$.

- Another perspective: Laurent series with a finite negative part correspond to the field of fractions of power series. Specifically, a power series $a_0 + a_1 z + \dots$ has an inverse in $\mathbb{C}[[z]]$ if and only if $a_0 \neq 0$, and otherwise $(a_n z^n + \dots)^{-1} = \frac{1}{a_n z^n}(1 + \dots)$. Therefore, a ratio of two non-trivial power series gives a Laurent series, which defines a meromorphic function.
- The converse is also true (though we will not prove it here): every meromorphic function is the quotient of two analytic functions. Thus, the meromorphic functions form the field of fractions of the ring of analytic functions.

Assume f is meromorphic on all of $\mathbb{C}$ (i.e., f is analytic on $\mathbb{C} \setminus \{p_i\}$, with poles at p_i). If $|f(z)|$ is either bounded or tends to infinity as $|z| \to \infty$, then the function $g(w) = f\left(\frac{1}{w}\right)$ has a removable singularity or a pole at $w = 0$, so it is meromorphic near 0. This implies that $\hat{f}$ can be extended to the Riemann sphere by setting $\hat{f}(\infty) = \hat{g}(0)$. Therefore, if $f(z)$ and $f\left(\frac{1}{z}\right)$ are meromorphic, we can extend f to an analytic function $\hat{f} : S \to S$ on the entire Riemann sphere.

In fact, such an extension $\hat{f}$ is necessarily a rational function. Indeed, we have the following result:

Theorem 9.43. *If f is an entire function (i.e., analytic on all of $\mathbb{C}$) and satisfies $|f(z)| \leq M|z|^n$ for some constants $M, n > 0$ as $|z| \to \infty$, then f is a polynomial of degree at most n.*

This follows from Cauchy's bound for derivatives: $f^{(n)}(z)$ is a bounded entire function and hence must be constant.

Corollary 9.44. *If $f : S \to S$ is analytic (i.e., both $f(z)$ and $f\left(\frac{1}{z}\right)$ are meromorphic), then f is a rational function.*

Proof. The fact that $g(w) = f(w)f\left(\frac{1}{w}\right)$ is meromorphic near $w = 0$ gives a bound of the form $|g(w)| \leq \frac{c}{|w|^n}$ as $w \to 0$, implying that $|f(z)| \leq c|z|^n$ for $z \in \mathbb{C}$ as $z \to \infty$.

Although f is not an entire function (it has poles), it has only finitely many poles. The poles of f correspond to the zeros of $\frac{1}{f}$, which are isolated, and since S is compact, the set of poles of f is finite.

Thus, there exists a polynomial $P(z) = \prod (z - p_i)^{n_i}$, where p_i are the poles of f and n_i are their orders, such that $P(z)f(z)$ extends to an entire function on $\mathbb{C}$, and this function satisfies the bound $|P(z)f(z)| \leq C'|z|^{n+\deg P}$ as $z \to \infty$. By the previous theorem, $P(z)f(z)$ must be a polynomial.

□

9.9 Local Behavior of Analytic Functions

Cauchy's integral formula can be viewed as a mean value theorem.

Theorem 9.45. *If f is analytic on $U \supset \overline{B_r(z)}$, then $f(z)$ is the average value of f on the circle $S^1(z,r)$.*

Proof. By Cauchy's integral formula, we have:

$$f(z) = \frac{1}{2\pi i}\int_{S^1(z,r)} \frac{f(w)}{w-z}\,dw = \frac{1}{2\pi i}\int_0^{2\pi} \frac{f(z+re^{i\theta})}{re^{i\theta}}\,d(re^{i\theta}) = \frac{1}{2\pi}\int_0^{2\pi} f(z+re^{i\theta})\,d\theta.$$

□

Theorem 9.46 (The Maximum Principle)**.** *If f is analytic on an open, connected set $U \subset \mathbb{C}$ and non-constant, then $|f|$ does not achieve its maximum value anywhere in U. In particular, if f is analytic on U and continuous on $\overline{U}$, where $\overline{U}$ is compact, then the maximum of $|f|$ on $\overline{U}$ is achieved on the boundary of U.*

Proof. Let $z_0 \in U$ and let $r > 0$ be small enough so that $\overline{B}_r(z_0) \subset U$. Then,

$$|f(z_0)| = \left|\frac{1}{2\pi}\int_0^{2\pi} f(z_0+re^{i\theta})\,d\theta\right| \le \frac{1}{2\pi}\int_0^{2\pi} |f(z_0+re^{i\theta})|\,d\theta \le \max_{S^1(z_0,r)} |f|.$$

If $|f|$ has a local maximum at z_0, then $\max_{S^1(z_0,r)} |f| = |f(z_0)|$, and the inequalities above become equalities. This implies that $|f(z)| = |f(z_0)|$ for all $z \in S^1(z_0,r)$.

In fact, $f(z) = f(z_0)$: if $\arg(f(z))$ varies, then the first inequality becomes strict (for instance, rescale so that $f(z_0) = 1$). In this case, $|f(z)| \le 1$, so $\mathrm{Re}(f(z)) \le 1$ for all $z \in S^1(z_0,r)$, and equality implies that $\mathrm{Re}(f(z)) = 1$ for all $z \in S^1(z_0,r)$. Since $|f(z)| \le 1$, we conclude that $f(z) = 1$ for all $z \in S^1(z_0,r)$.

Since f is analytic, we have $f(z) - f(z_0) = 0$ for all $z \in S^1(z_0,r)$, which implies that the zeros of $f(z) - f(z_0)$ are not isolated (as the zeros of nontrivial analytic functions are isolated). Thus, $f(z) - f(z_0) = 0$ on U, and hence f is constant on U. □

Remark 9.47. *This also implies the maximum principle for $Re(f)$, since $|e^f| = e^{Re(f)}$ has no local maximum.*

One nice (non-local) consequence is a contraction principle:

Theorem 9.48 (The Schwarz Lemma)**.** *Let f be analytic on the unit disk $D = \{z : |z| < 1\}$, and suppose $|f(z)| < 1$ for all $z \in D$ (i.e., $f : D \to D$), and $f(0) = 0$. Then:*

- $|f'(0)| \le 1$.

- $|f(z)| < |z|$ *for all* $z \in D \setminus \{0\}$.

Moreover, if equality holds in either of these inequalities, then $f(z) = e^{i\theta}z$ *for some* $e^{i\theta} \in S^1$.

Proof. Write $f(z) = \sum_{n=1}^{\infty} a_n z^n = zF(z)$, where $F(z) = \sum_{n=0}^{\infty} a_{n+1} z^n$ is analytic. For $|z| = r \in (0,1)$, we have $|F(z)| = \left|\frac{f(z)}{z}\right| \leq \frac{1}{r}$. Hence, by the maximum principle, $|F(z)| \leq \frac{1}{r}$ whenever $|z| \leq r$. Taking $r \to 1$, we get $|F(z)| \leq 1$ for all $z \in D$.

Thus, the bound on $f'(0) = F(0)$ follows. Moreover, if $|F(z)| = 1$ is achieved anywhere inside D, then F must be constant, and hence $f(z) = e^{i\theta}z$. □

Remark 9.49.

- *The bound on* $|f'(0)|$ *is the same as the bound one obtains from Cauchy's integral formula. The Schwarz Lemma is a strengthening of this result, providing a pointwise bound* $|f(z)| \leq |z|$ *globally on the disk.*
- *By composing* f *with fractional linear transformations, we can obtain Schwarz-type bounds in various other situations, such as when* f *maps a disk to a half-plane, etc.*

We can also deduce a stronger local result:

Theorem 9.50 (The Open Mapping Principle)**.** *A non-constant analytic function is an open mapping, i.e., if* U *is open, then* $f(U)$ *is open.*

In other words, if f is non-constant and analytic at z_0, then for all $r > 0$, there exists $\epsilon > 0$ such that $f(B_r(z_0)) \supset B_\epsilon(f(z_0))$. This implies that $|f(z)|$, $\mathrm{Re}(f(z))$, and similar functions cannot have a local maximum.

First, we prove:

Proposition 9.51. *If* $f(z)$ *has an isolated zero at* $z = z_0$*, then there exists an analytic function* g *defined near* z_0*, with* $g(z_0) = 0$*,* $g'(z_0) \neq 0$*, and* $n \geq 1$ *such that* $f(z) = g(z)^n$.

Proof. Let n be the order of the zero of f, i.e., write $f(z) = \sum_{k=n}^{\infty} a_k (z - z_0)^k = a_n(z - z_0)^n(1 + h(z))$ with $h(z_0) = 0$. There exists a neighborhood V of z_0 such that $|h(z)| < 1$ for all $z \in V$. Over V, we can define $g(z) = a_n^{1/n}(z - z_0)(1 + h(z))^{1/n}$, where $(1 + h(z))^{1/n} = \exp\left(\frac{1}{n}\log(1 + h(z))\right)$ is well-defined for $|h(z)| < 1$. □

Now, we prove the following theorem:

Theorem 9.52. *For* $z_0 \in U$*, if* $f(z) - f(z_0) = g(z)^n$ *for some* $n \geq 1$ *and* $g(z_0) = 0$*,* $g'(z_0) \neq 0$*, then by the inverse function theorem,* g *is a local diffeomorphism at* z_0 *(since* $g'(z_0) \neq 0$*), hence an open mapping near* z_0*. This implies that*

there exists an open neighborhood $V \subset U$ such that $g(V) \ni 0$ contains some ball $B_\epsilon(0)$, and hence, taking the n^{th} power, we have $f(V) \supset B(f(z_0), \epsilon^n)$.

9.10 Harmonic Functions

Another important class of functions that satisfy the mean value and maximum principles is the class of harmonic functions:

Definition 9.53. *A C^2 function $f : U(\subset \mathbb{R}^n) \to \mathbb{R}$ is **harmonic** if $\Delta f = \sum_{i=1}^n \frac{\partial^2 f}{\partial x_i^2} = 0$.*

Remark 9.54. *This is of significant physical importance! For example, electric and gravitational potentials in a vacuum are harmonic, as is the temperature distribution at thermal equilibrium, and so on.*

Real analysis provides general methods for studying harmonic functions, but in the case of two real variables $f(x, y)$, the situation is closely related to complex analysis.

Proposition 9.55. *Let $u : U \subset \mathbb{C} \to \mathbb{R}$ be a C^2 function. Then u is harmonic if $\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 4\frac{\partial^2 u}{\partial z \partial \overline{z}} = 0$.*

Proof. Since $\frac{\partial^2 u}{\partial x \partial y} = \frac{\partial^2 u}{\partial y \partial x}$, we have

$$\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right)\left(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right)u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}.$$

□

Theorem 9.56. *If $f = u + iv$ is analytic, then $u = Re\ f$ and $v = Im\ f$ are harmonic.*

Proof. Two proofs:

1. Cauchy-Riemann equations: $\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}$ and $\frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$, we can compute
$$\Delta u = \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial u}{\partial y}\right) = \frac{\partial}{\partial x}\left(\frac{\partial v}{\partial y}\right) - \frac{\partial}{\partial y}\left(\frac{\partial v}{\partial x}\right) = 0.$$
2. Complex variables: Since $u = \frac{1}{2}\left(f + \overline{f}\right)$, we have $\Delta f = 4\frac{\partial}{\partial z}\left(\frac{\partial f}{\partial \overline{z}}\right) = 0$, and similarly, $\Delta(\overline{f}) = 4\frac{\partial}{\partial \overline{z}}\left(\frac{\partial \overline{f}}{\partial z}\right) = 0$.

□

What is unique about harmonic functions in two variables is that there is a converse:

Theorem 9.57. *If u is harmonic on a simply-connected open set $U \subset \mathbb{C}$, then there exists an analytic function $f : U \to \mathbb{C}$ such that $u = Re\, f$, i.e., there exists a harmonic function $v : U \to \mathbb{R}$ such that $u + iv$ is analytic.*

Example 9.58. *Consider $u = \log|z| = Re(\log z)$ on a domain that does not enclose the origin. Here, $v = arg(z)$. This example demonstrates that the assumption on U is necessary, as v is not single-valued on C^*.*

Proof. Given that u is harmonic, define the complex-valued 1-form $\alpha = 2\frac{\partial u}{\partial z}\, dz = \left(\frac{\partial u}{\partial x} - i\frac{\partial u}{\partial y}\right)(dx + idy)$. Then α is closed because $d\alpha = \frac{2\partial}{\partial \overline{z}}\left(\frac{\partial u}{\partial z}\right) d\overline{z}\, dz$ and $2\frac{\partial^2 u}{\partial \overline{z}\partial z} = \frac{1}{2}\Delta u = 0$. In real differential terms, $\mathrm{Re}(\alpha) = \frac{\partial u}{\partial x}\, dx + \frac{\partial u}{\partial y}\, dy = du$, so $\mathrm{Re}(\alpha)$ is exact and hence closed. Similarly, $\mathrm{Im}(\alpha) = \frac{\partial u}{\partial x}\, dy - \frac{\partial u}{\partial y}\, dx$ is closed, using the fact that $\Delta u = 0$.

Since U is simply connected, closed 1-forms on U are exact. Therefore, there exists a function $f : U \to \mathbb{C}$ such that $df = \alpha$. We can construct f by integration: $f(z) = \int_{z_0}^{z} \alpha$, which is path-independent by Stokes' theorem since $d\alpha = 0$ and U is simply connected.

Now, $df = \frac{\partial f}{\partial z}\, dz + \frac{\partial f}{\partial \overline{z}}\, d\overline{z} = 2\frac{\partial u}{\partial z}\, dz$, which implies that $\frac{\partial f}{\partial \overline{z}} = 0$, i.e., f is analytic. Since $d(\mathrm{Re} f) = \mathrm{Re}(\alpha) = du$, we can ensure that $\mathrm{Re}(f) = u$ by adding a constant. □

Now that we know harmonic functions are secretly parts of analytic functions, we obtain the following corollary:

Corollary 9.59.
- *Any C^2 harmonic function is actually C^∞.*
- *Harmonic functions satisfy the mean value theorem:*

$$u(z) = \frac{1}{2\pi}\int_0^{2\pi} u(z + re^{i\theta})\, d\theta.$$

Another pair of deep results (which we won't prove here) concern the existence of analytic mappings and harmonic functions.

Theorem 9.60 (The Riemann Mapping Theorem)**.** *If $U \subset \mathbb{C}$ is a non-empty simply connected open subset with $U \neq \mathbb{C}$, then there exists a* ***biholomorphism*** *$\varphi : U \xrightarrow{\sim} D = \{|z| < 1\}$, i.e., an analytic bijection with an analytic inverse.*

Example 9.61. *Can you find an explicit biholomorphism between a quarter disk, half disk, $(-\infty, 0) \times (0, 1)$, or the unit disk and the upper half-plane $\mathbb{R} \times (0, 1)$?*

The existence of solutions to Dirichlet's problem (harmonic functions with prescribed values at the boundary of a domain) can be thought of as an analogue for harmonic functions.

Theorem 9.62. *If $U \subset \mathbb{C}$ is a simply connected bounded open set with a sufficiently nice boundary (e.g., ∂U is piecewise smooth), and $f \in C^0(\partial U, \mathbb{R})$ is any continuous function, then there exists a unique function $u \in C^0(\overline{U}, \mathbb{R})$ such that $u|_{\partial U} = f$ and u is harmonic inside U.*

Remark 9.63. *Uniqueness follows easily from the maximum principle: if $u - v = 0$ on ∂U and $u - v$ is harmonic, then $u - v = 0$ everywhere in U.*

One way to prove this theorem is to first establish it for the unit disk, using Fourier series to reduce to trigonometric polynomials; $\sum c_n e^{in\theta} \to \sum_{n \geq 0} c_n z^n + \sum_{n>0} c_n \overline{z}^{|n|}$. Then, we use the Riemann mapping theorem to map $U \xrightarrow{\varphi \sim} D$, where u is harmonic if and only if $u \circ \varphi$ is.

9.11 Open Mapping Principle

Consider analytic functions. There is a stronger local result known as the open mapping theorem.

Theorem 9.64 (The Open Mapping Theorem)**.** *If f is a nonconstant analytic function, then f is an open mapping. That is, if U is open, then $f(U)$ is open.*

In other words, if f is nonconstant and analytic at z_0, then for every $r > 0$, there exists an $\epsilon > 0$ such that

$$f(B_r(z_0)) \supset B_\epsilon(f(z_0)),$$

where $B_r(z_0)$ denotes the ball of radius r around z_0, and $B_\epsilon(f(z_0))$ denotes the ball of radius ϵ around $f(z_0)$. This implies that $|f(z)|$, $\mathrm{Re}(f(z))$, and other similar quantities cannot have a local maximum.

We first prove the following result:

Proposition 9.65. *If $f(z)$ has an isolated zero at z_0, then there exists an analytic function g defined near z_0 such that $g(z_0) = 0$, $g'(z_0) \neq 0$, and for some $n \geq 1$, we have $f(z) = g(z)^n$.*

Proof. Let n be the order of the zero of f at z_0. We can express $f(z)$ as

$$f(z) = \sum_{k=n}^{\infty} a_k (z - z_0)^k = a_n (z - z_0)^n (1 + h(z)),$$

where $h(z)$ is analytic, and $h(z_0) = 0$. Furthermore, there exists a neighborhood V of z_0 such that $|h(z)| < 1$ for all $z \in V$. Over this neighborhood V, we can define the function $g(z)$ as

$$g(z) = a_n^{1/n} (z - z_0) \left(1 + h(z)\right)^{1/n},$$

where the expression $(1 + h(z))^{1/n}$ is well-defined for $|h(z)| < 1$, and it can be written as

$$(1 + h(z))^{1/n} = \exp\left(\frac{1}{n} \log(1 + h(z))\right),$$

which is analytic on V.

Thus, we have $f(z) = g(z)^n$, and the function $g(z)$ satisfies the required conditions.

□

Now, we can proceed to prove the open mapping theorem.

Proof. Let $z_0 \in U$, where U is open. From the previous proposition, we know that there exists a function $g(z)$ such that $f(z) - f(z_0) = g(z)^n$ for some $n \geq 1$, with $g(z_0) = 0$ and $g'(z_0) \neq 0$. By the inverse function theorem, since $g'(z_0) \neq 0$, the function g is a local diffeomorphism at z_0. This implies that g is an open mapping near z_0, and that g has a continuous, in fact analytic, inverse in some neighborhood of z_0.

Hence, for every open set V containing z_0 (which is contained in the domain of g), the image $g(V)$ contains a ball $B_\epsilon(0)$ centered at 0. Consequently, for the function f, we have

$$f(V) = g(V)^n \supset B(f(z_0), \epsilon^n),$$

which shows that $f(V)$ contains an open set around $f(z_0)$.

Thus, f is an open mapping, completing the proof.

□

10 Complex Analysis II

10.1 Residue Calculus

Instead of using Cauchy's integral formula to study the behavior of analytic functions, we now use it to evaluate integrals.

Assume we want to evaluate the integral $\int_\gamma f(z)\,dz$, where $\gamma = \partial D$ and f is analytic in a neighborhood $U \supset \overline{D} - \{p_1, \ldots, p_n\}$ (or later, a definite integral whose value can be related to $\int_\gamma$).

Definition 10.1. *The **residue** of a function f at a point p is given by*

$$Res_p(f) = \frac{1}{2\pi i} \int_{S^1(p,\epsilon)} f(z)\,dz,$$

where $\epsilon > 0$ is small enough so that f is analytic in the punctured disk $D^*(p,\epsilon) = D(p,\epsilon) - \{p\}$.

If we express f as a Laurent series

$$f(z) = \sum_{n=-\infty}^{\infty} a_n (z-p)^n$$

in the region $D^*(p,\epsilon)$, then the residue of f at p is the coefficient of the -1 term in this Laurent series, i.e.,

$$\mathrm{Res}_p(f) = a_{-1}.$$

Thus, the residue is easiest to calculate if f has a simple pole (i.e., pole of order 1) at p. In this case, the residue is given by

$$\mathrm{Res}_p(f) = \lim_{z \to p} (z-p) f(z).$$

Otherwise, it may be necessary to compute the residue by determining part of the Laurent series for f. For example, in the case of rational functions, partial fraction decomposition can help achieve this.

Now, applying Cauchy's theorem to the domain $D - \bigcup D(p,\epsilon)$, we obtain the following result.

Theorem 10.2 (Residue Theorem)**.** *Let $\overline{D}$ be a compact domain with piecewise smooth boundary $\gamma = \partial D$, and let $P \subset int(D)$ be a finite set of points. If f is analytic on $U \supset \overline{D} - P$, then*

$$\frac{1}{2\pi i} \int_\gamma f(z)\,dz = \sum_{p \in P} Res_p(f).$$

We now explore how to use this theorem to evaluate various definite integrals.

Example 10.3. *Evaluate the integral* $\int_0^{2\pi} R(\sin\theta, \cos\theta)\, d\theta$ *(or* $R(e^{i\theta})$*), where* R *is a rational function with no poles on* S^1*. For instance, consider the integral*

$$\int_0^{2\pi} \frac{d\theta}{a + \cos\theta}, \quad a > 1.$$

To convert this into a path integral on the unit circle, let $z = e^{i\theta}$*, so that* $d\theta = \frac{dz}{iz}$ *and* $\cos\theta = \frac{z+z^{-1}}{2}$*. This gives:*

$$\int_0^{2\pi} \frac{d\theta}{a + \cos\theta} = \int_{S^1} \frac{dz}{i\,(z^2 + 2az + 1)}.$$

The poles of the integrand are at $p_\pm = -a \pm \sqrt{a^2 - 1}$*, and only* $p_+ = -a + \sqrt{a^2 - 1}$ *lies inside the unit circle.*

We calculate the residue at p_+*. Using partial fractions, we express the integrand as*

$$f(z) = \frac{1}{(z - p_+)(z - p_-)} \left(\frac{1}{z - p_+} - \frac{1}{z - p_-} \right).$$

The residue at p_+ *is given by*

$$Res_{p_+}(f) = \frac{1}{2\sqrt{a^2 - 1}}.$$

Since this is a simple pole, we also find that

$$Res_{p_+}(f) = \lim_{z \to p_+} (z - p_+) f(z) = \frac{1}{p_+ - p_-} = \frac{1}{2\sqrt{a^2 - 1}}.$$

Thus, the value of the integral is

$$\int_0^{2\pi} \frac{d\theta}{a + \cos\theta} = 4\pi Res_{p_+}(f) = \frac{2\pi}{\sqrt{a^2 - 1}}.$$

Example 10.4. *Consider the integral* $\int_{-\infty}^{\infty} f(x)\, dx$*, where* f *is a rational function* $\frac{P(x)}{Q(x)}$*. Assume that* $Q(x)$ *has no real roots and that* $\deg Q \geq \deg P + 2$*, so the integral converges.*

We use the fact that

$$\int_{-\infty}^{\infty} f(x)\, dx = \lim_{R \to \infty} \int_{-R}^{R} f(x)\, dx,$$

and we complete the segment $[-R, R]$ *to a closed curve in the upper half-plane by adding a semicircle of radius* R*. This gives*

$$\int_{-R}^{R} f(x)\, dx + \int_{C_R} f(z)\, dz = 2\pi i \sum_{Im(p) > 0, |p| < R} Res_p(f).$$

Now, since $f = \frac{P}{Q}$ *with* $\deg Q \geq \deg P + 2$*, we have* $|f(z)| \leq \frac{c}{|z|^2}$*, so*

$$\lim_{R\to\infty} \int_{C_R} f(z)\, dz = 0.$$

Hence, as $R \to \infty$*, we get*

$$\int_{-\infty}^{\infty} f(x)\, dx = 2\pi i \sum_{Im(p)>0} Res_p(f).$$

We can use $\lim_{z\to p}(z-p)f(z)$ *to compute the residues if all poles are simple; otherwise, partial fractions can be used.*

For example, consider the integral

$$\int_{-\infty}^{\infty} \frac{dx}{x^2+1}.$$

Using the residue theorem, we get

$$\int_{-\infty}^{\infty} \frac{dx}{x^2+1} = 2\pi i Res_{z=i}\left(\frac{1}{z^2+1}\right) = \pi.$$

Example 10.5. *Consider the integral of a mixed rational and exponential function:*

$$\int_{-\infty}^{\infty} \frac{e^{iz}}{1+z^2}\, dz.$$

We apply the residue theorem by closing the contour in the upper half-plane. Since $|e^{iz}| = e^{-Im(z)} \leq 1$ *in the upper half-plane, the integral along the semicircle tends to* 0 *as the radius increases.*

The residue at $z = i$ *is computed as follows:*

$$Res_{z=i}\left(\frac{e^{iz}}{1+z^2}\right) = e^{-1} Res_{z=i}\left(\frac{1}{1+z^2}\right) = \frac{1}{2ie}.$$

Thus, the integral evaluates to

$$\int_{-\infty}^{\infty} \frac{e^{iz}}{1+z^2}\, dz = \frac{\pi}{e}.$$

By separating the real and imaginary parts, we obtain:

$$\int_{-\infty}^{\infty} \frac{\cos x}{1+x^2}\, dx = \frac{\pi}{e}, \quad \int_{-\infty}^{\infty} \frac{\sin x}{1+x^2}\, dx = 0.$$

Example 10.6. *Next, consider the integral*

$$\int_{-\infty}^{\infty} \frac{\sin x}{x}\, dx,$$

which appears in G.H. Hardy's note in the Mathematical Gazette (1909), where various methods for evaluating it are discussed. This integral converges, though not absolutely. The function $\frac{\sin z}{z} = 1 - \frac{z^2}{3!} + \frac{z^4}{5!} + \cdots$ *is analytic in the entire complex plane, so there are no residues to compute. However, when expressed as* $\frac{\sin z}{z} = \frac{e^{iz} - e^{-iz}}{2iz}$*, the function tends to infinity both in the upper and lower half-planes. Thus, we cannot simply close the contour to a half-disc in the usual way.*

Nevertheless, we can proceed by writing $\frac{\sin x}{x}$ *as the limit*

$$\frac{\sin x}{x} = \lim_{a \to \infty} \frac{x \sin x}{a^2 + x^2},$$

and, after a careful discussion of the convergence as $a \to \infty$ *and the interchange of limits, we find that taking* $a \to 0$ *is legitimate.*

It is more instructive, however, to adjust the previous argument to handle $a = 0$*. The issue lies in the fact that for* $x \in \mathbb{R}$*, we have* $\frac{\sin x}{x} = Im\left(\frac{e^{ix}}{x}\right)$*, but* $\frac{e^{iz}}{z}$ *has a pole at* $z = 0$*, which lies on the path of integration. In fact, the integral*

$$\int_0^\infty \frac{e^{ix}}{x}\, dx$$

is divergent at 0*.*

The solution is to modify the contour of integration to avoid 0 *by carving out a small disc from the large semicircle on the rectangle. Specifically, we have the following steps:*

- *First, we write the integral as*

$$\int_{-\infty}^{\infty} \frac{\sin x}{x}\, dx = \lim_{R \to \infty, \epsilon \to 0} \int_{[-R,-\epsilon] \cup [\epsilon, R]} \frac{\sin x}{x}\, dx = \lim_{R \to \infty, \epsilon \to 0} Im \int_{[-R,-\epsilon] \cup [\epsilon, R]} \frac{e^{iz}}{z}\, dz.$$

- *The integral along the boundary of the small disc* $D_{R,\epsilon}$ *satisfies*

$$\int_{\partial D_{R,\epsilon}} \frac{e^{iz}}{z}\, dz = 0$$

  by Cauchy's theorem, as there are no poles inside the region $D_{R,\epsilon}$*.*

- *The integral on the semicircle of radius* R *tends to* 0 *as* $R \to \infty$*, as before. This follows from the fact that*

$$\left| \frac{e^{iz}}{z} \right| = \frac{e^{-Im(z)}}{R},$$

  which decays exponentially, and we can separate the regions where $Im(z) < A$ *and* $Im(z) > A$ *for large* A*.*

- *On the small semicircle of radius ϵ, the residue at $z = 0$ is*

$$Res_0\left(\frac{e^{iz}}{z}\right) = 1.$$

Thus, we can write $\frac{e^{iz}}{z} = \frac{1}{z} + g(z)$, where $g(z)$ is analytic near $z = 0$ (specifically, $g(z) = \frac{e^{iz}-1}{z}$). Since $g(z)$ is bounded, we have

$$\int_{C_\epsilon} g(z)\,dz \to 0 \quad as \quad \epsilon \to 0,$$

whereas

$$\int_{C_\epsilon} \frac{1}{z}\,dz = i\pi.$$

Combining these results, we find that

$$\lim_{\epsilon\to 0, R\to\infty} \int_{[-R,-\epsilon]\cup[\epsilon,R]} \frac{e^{iz}}{z}\,dz = i\pi.$$

Thus, we conclude

$$\int_{-\infty}^{\infty} \frac{\sin x}{x}\,dx = \pi.$$

Example 10.7. *Finally, consider the class of integrals involving non-integer powers of z. For example, we consider the integral*

$$I(\alpha) = \int_0^\infty \frac{x^\alpha}{1+x^2}\,dx,$$

for $0 < \alpha < 1$, which converges at $x = \infty$. If $\alpha = \frac{p}{q} \in \mathbb{Q}$, we can evaluate the integral by substitution $x = u^1$, which transforms the integral into a rational function. However, for a more general approach, we proceed as follows.

The function $\frac{z^\alpha}{1+z^2}$ is not single-valued and analytic on the entire complex plane, so we must be cautious when using residues.

We proceed by using a "keyhole" contour, where the region of integration is $0 \le |z| \le R$ with a slit along the positive real axis. In this case, the two portions along $[\epsilon, R]$ do not cancel out, since the values of $\frac{z^\alpha}{1+z^2}$ on either side of the real axis are not the same.

To handle this, we define

$$z^\alpha = e^{\alpha \log z},$$

where $Im(\log z) \in (0, 2\pi)$. When going around the origin, $\log x \to \log x + 2\pi i$, so x^α is multiplied by $e^{2\pi i\alpha}$. Thus, the contour integral is

$$\int_{\partial D} \frac{z^\alpha}{1+z^2}\,dz = \int_\epsilon^R \frac{x^\alpha}{1+x^2}\,dx + \int_{C_R} \frac{z^\alpha}{1+z^2}\,dz - \int_\epsilon^R \frac{e^{2\pi i\alpha}x^\alpha}{1+x^2}\,dx - \int_{C_\epsilon} \frac{z^\alpha}{1+z^2}\,dz.$$

The integrals on the large and small semicircles C_R and C_ϵ tend to zero as $R \to \infty$ and $\epsilon \to 0$. Thus, we have

$$\lim_{\epsilon \to 0, R \to \infty} \int_{\partial D} \frac{z^\alpha}{1+z^2}\, dz = (1 - e^{2\pi i \alpha}) \int_0^\infty \frac{x^\alpha}{1+x^2}\, dx.$$

On the other hand, applying the residue theorem gives

$$\int_{\partial D} \frac{z^\alpha}{1+z^2}\, dz = 2\pi i \left(Res_{z=i} \left(\frac{z^\alpha}{1+z^2} \right) + Res_{z=-i} \left(\frac{z^\alpha}{1+z^2} \right) \right).$$

We compute the residues as follows:

- *At $z = i$, we have*

$$Res_{z=i} \left(\frac{z^\alpha}{1+z^2} \right) = \frac{1}{2i} e^{\alpha \log i} = \frac{1}{2i} e^{i\frac{\pi}{2}\alpha}.$$

- *At $z = -i$, we similarly get*

$$Res_{z=-i} \left(\frac{z^\alpha}{1+z^2} \right) = -\frac{1}{2i} e^{3i\frac{\pi}{2}\alpha}.$$

Therefore, we conclude that

$$\int_0^\infty \frac{x^\alpha}{1+x^2}\, dx = \pi \frac{e^{i\pi\alpha/2} - e^{3i\pi\alpha/2}}{1 - e^{2\pi i \alpha}} = \frac{\pi \sin\left(\frac{\pi\alpha}{2}\right)}{\sin(\pi\alpha)} = \frac{\pi}{2\cos\left(\frac{\pi\alpha}{2}\right)}.$$

10.2 Infinite Sum and Product Expansions

We have seen that if f is analytic in the annulus $\{R_1 < |z| < R_2\}$, then it has a Laurent series expansion

$$f(z) = \sum_{n=-\infty}^{\infty} a_n z^n,$$

which may or may not have a finite negative part.

If the inner radius is $R_1 = 0$, then a finite negative part corresponds to a pole at $z = 0$, while an infinite negative part corresponds to an essential singularity. However, if $R_1 > 0$, this is not necessarily the case!

Example 10.8. *Consider the function $\frac{1}{1-z}$, which has a pole at $z = 1$. We can express this function using two different Laurent series:*

- *For $|z| < 1$, we have*

$$\frac{1}{1-z} = 1 + z + z^2 + \cdots \quad (R_2 = 1).$$

- *For $|z| > 1$, we can write*

$$\frac{1}{1-z} = \frac{-1}{z\left(1-\frac{1}{z}\right)} = -z^{-1} - z^{-2} - z^{-3} - \cdots \quad (R_1 = 1).$$

OIn this example, and for most rational functions, Laurent series are not the best representation. Instead, partial fractions or product expansions are more suitable.

- **Products:** If $R(z) = \frac{P(z)}{Q(z)}$, we can factor $R(z) = c\frac{\prod_{i=1}^{k}(z-a_i)^{n_i}}{\prod_{i=1}^{l}(z-b_i)^{m_i}}$.
- **Sums (partial fractions):** If the poles are simple, we can write

$$R(z) = \frac{c_1}{z-b_1} + \cdots + \frac{c_l}{z-b_l} + S(z),$$

  where $c_i \in \mathbb{C}$ and $S(z)$ is a polynomial. More generally, we can express

$$R(z) = \frac{C_1(z)}{(z-b_1)^{m_1}} + \cdots + \frac{C_l(z)}{(z-b_l)^{m_l}} + S(z),$$

  where $C_1(z), \ldots, C_l(z)$ are polynomials with $\deg(C_i) \leq m_i - 1$.

We will explore how to find similar infinite sum or product expansions for general meromorphic functions.

Starting point: If $f(z)$ is meromorphic with a pole of order m at $b \in \mathbb{C}$, then we can write

$$f(z) = \frac{g(z)}{(z-b)^m},$$

where $g(z)$ is analytic in a neighborhood of b. Expanding $g(z)$ as a power series in $(z-b)$, we have

$$g(z) = \sum_{n=0}^{\infty} a_n (z-b)^n.$$

Thus, the Laurent series for f has a finite negative part, as shown earlier:

$$f(z) = \left[\frac{a_0}{(z-b)^m} + \frac{a_1}{(z-b)^{m-1}} + \cdots + \frac{a_{m-1}}{z-b}\right] + h(z),$$

where the bracketed portion is the polar part of f at $z = b$, and $h(z) = \sum_{n=0}^{\infty} a_{m+n}(z-b)^n$ is analytic near b.

This resembles partial fractions, and in fact, for rational functions, it is partial fractions. If f is meromorphic with finitely many poles $b_1, \ldots, b_l$, by induction on the number of poles (note that the remainder $h(z)$ has one fewer pole than f), we obtain

$$f(z) = \frac{P(z)}{Q(z)} \implies \frac{c_1(z)}{(z-b_1)^{m_1}} + \cdots + \frac{c_l(z)}{(z-b_l)^{m_l}} + g(z),$$

where $C_i(z)$ are polynomials of degree less than m_i, and $g(z)$ is analytic everywhere.

What if there are infinitely many poles?

Given that $f(z)$ is meromorphic on all of $\mathbb{C}$, with infinitely many isolated poles $b_1, b_2, \ldots$, we consider the polar part near each b_j, which takes the form

$$P_j\left(\frac{1}{z-b_j}\right) = \frac{a_{-m}}{(z-b_j)^m} + \cdots + \frac{a_{-1}}{z-b_j},$$

a polynomial without a constant term in the variable $\frac{1}{z-b_j}$. We hope to express

$$f(z) = \sum_{j=1}^{\infty} P_j\left(\frac{1}{z-b_j}\right) + g(z),$$

where $g(z)$ is an entire function, i.e., it has no poles.

Problem 10.9.

- *When do these sums converge? Can they converge uniformly?*
- *What meromorphic functions can be represented in such a way?*
- ***Existence:*** *Given a discrete set of poles b_j and orders m_j, does there exist a meromorphic function with exactly these poles? Can we prescribe the polar parts $P_j\left(\frac{1}{z-b_j}\right)$ arbitrarily?*

An apparent problem: the expression $\sum_{n\in\mathbb{Z}} \frac{1}{z-n}$ does not seem to make sense.

Example 10.10. *Consider the function $f(z) = \frac{\pi^2}{\sin^2(\pi z)}$, which has poles (of order 2) exactly at the integers.*

The polar part at $z = 0$ can be found by expanding $\sin(\pi z) = \pi z - \frac{\pi^3}{6}z^3 + \cdots$, which gives

$$\begin{aligned}\sin^2(\pi z) &= \pi^2 z^2 - \frac{\pi^4}{3}z^4 + \cdots \\ &= \pi^2 z^2\left(1 - \frac{\pi^2}{3}z^2 + \cdots\right).\end{aligned}$$

Thus, we have

$$\frac{\pi^2}{\sin^2(\pi z)} = \frac{1}{z^2}\left(1 + \frac{\pi^2}{3}z^2 + \cdots\right),$$

which implies the polar part at $z = 0$ is simply $\frac{1}{z^2}$. Since f is periodic ($f(z+1) = f(z)$), the polar part at $z = n \in \mathbb{Z}$ is $\frac{1}{(z-n)^2}$.

Now, consider the sum

$$h(z) = \sum_{n\in\mathbb{Z}} \frac{1}{(z-n)^2},$$

which is convergent for all $z \in \mathbb{C}\setminus\mathbb{Z}$, and the convergence is uniform on compact subsets of $\mathbb{C} \setminus \mathbb{Z}$ (this can be proven using the M-test). The sum defines an analytic function on $\mathbb{C} \setminus \mathbb{Z}$, which can be checked to have the correct behavior (pole of order 2 with polar part $\frac{1}{(z-n)^2}$ at each $n \in \mathbb{Z}$).

Hence, we can write

$$\frac{\pi^2}{\sin^2(\pi z)} = \sum_{n\in\mathbb{Z}} \frac{1}{(z-n)^2} + g(z),$$

where $g(z)$ is an entire function. Since the polar parts cancel at each $z = n$, we conclude that $g(z)$ is an entire, periodic function: $g(z+1) = g(z)$. What is g?

Observe that for $Im(z) \to +\infty$, we have

$$|e^{i\pi z}| = e^{-\pi Im(z)} \ll e^{\pi Im(z)} = |e^{-i\pi z}|,$$

so

$$|f(z)| \approx \frac{4\pi^2}{e^{2\pi Im(z)}} \to 0 \quad as \quad Im(z) \to \pm\infty.$$

For $h(z)$, if $z = x + iy$ with $y \to +\infty$, we have

$$\left|\frac{1}{(z-n)^2}\right| = \frac{1}{|z-n|^2} = \frac{1}{(n-x)^2+y^2},$$

which implies the terms with $|n| < y$ are $\leq \frac{1}{y^2}$, and those with $|n| > y$ are $\leq \frac{1}{(n-1)^2}$, so

$$|h(z)| \leq \frac{C}{y}.$$

This shows that $g(z)$ is an entire function, periodic, and bounded, hence constant. Since $g(z) \to 0$ as $y \to \infty$, we conclude that $g(z) = 0$.

Thus, we find

$$\frac{\pi^2}{\sin^2(\pi z)} = \sum_{n\in\mathbb{Z}} \frac{1}{(z-n)^2}.$$

Problem 10.11. *Can we find a meromorphic function with simple poles at all integers and residue 1 at each? And can we express it as a partial fraction sum?*

The natural guess is

$$\sum_{n\in\mathbb{Z}} \frac{1}{z-n},$$

but this series does not converge.

Solution. To achieve convergence, we subtract the value of each term at $z = 0$, i.e., subtract $\frac{1}{n}$ from each term:

$$f(z) = \frac{1}{z} + \sum_{n\in\mathbb{Z}, n\neq 0} \left(\frac{1}{z-n} + \frac{1}{n}\right) = \frac{1}{z} + \sum_{n\neq 0} \frac{z}{n(z-n)}.$$

This series now converges for all $z \in \mathbb{C} \setminus \mathbb{Z}$, uniformly on compact subsets, and has the desired polar part at each integer. □

Problem 10.12. *Can we use a similar trick to build meromorphic functions with arbitrary poles and polar parts at each pole?*

Solution. The answer is yes, but we may need to add more complicated counterterms to ensure convergence. □

Theorem 10.13. *Let $\{b_j\}$ be an arbitrary set of complex numbers with no limit points, and let P_j be an arbitrary polynomial without constant term for each j. Then there exists a meromorphic function $f(z)$ on all of $\mathbb{C}$, analytic on $\mathbb{C}\setminus\{b_j\}$, and whose polar part at b_j is $P_j\left(\frac{1}{z-b_j}\right)$ for all j.*

Proof. The proof uses the same idea as above. To ensure convergence, we subtract from each $P_j\left(\frac{1}{z-b_j}\right)$ (for $b_j \neq 0$) a polynomial in z. Given $m_j \geq 0$ as an integer, let $q_j(z)$ be the sum of the terms of degree $\leq m_j$ in the Taylor series of $P_j\left(\frac{1}{z-b_j}\right)$ at $z = 0$. The point (see Ahlfors 5.2.1) is that we can choose the m_j's so that the series

$$f(z) = \sum_j \left(P_j\left(\frac{1}{z-b_j}\right) - q_j(z)\right)$$

converges on $\mathbb{C} \setminus \{b_j\}$.

How does one show this? First observe that if $\{b_j\}$ has no limit points, then the set $\{b_j\}$ is discrete and $|b_j| \to \infty$. Next, we need explicit bounds on the remainder $P_j\left(\frac{1}{z-b_j}\right) - q_j(z)$ from the Taylor series of $P_j\left(\frac{1}{z-b_j}\right)$.

For the base case, we have

$$\frac{1}{z-b_j} = -\frac{1}{b_j}\frac{1}{1-\frac{z}{b_j}} = -\frac{1}{b_j}\left(1 + \frac{z}{b_j} + \left(\frac{z}{b_j}\right)^2 + \cdots\right),$$

with remainder

$$\left(\frac{z}{b_j}\right)^{m_j+1} \frac{1}{z-b_j}.$$

Thus, we can estimate

$$\left|P_j\left(\frac{1}{z-b_j}\right) - q_j(z)\right| \leq \frac{1}{j^2}.$$

Since $|b_j| \to \infty$, this implies uniform convergence over compact subsets of $\mathbb{C}$, as all but finitely many terms of the series are bounded by $\sum \frac{1}{j^2}$.

□

Back to our function with simple poles at all integers:

$$f(z) = \frac{1}{z} + \sum_{n \neq 0} \left(\frac{1}{z-n} + \frac{1}{n} \right) = \frac{1}{z} + \sum_{n \neq 0} \frac{z}{n(z-n)}.$$

Since the series converges uniformly on compact subsets of $\mathbb{C} \setminus \mathbb{Z}$, we can differentiate term by term. Recall that if f_n is analytic and $f_n \to f$ uniformly, then $f_n' \to f'$ uniformly on compact subsets.

Thus, we find the derivative of $f(z)$:

$$f'(z) = -\sum_{n \in \mathbb{Z}} \frac{1}{(z-n)^2} = \frac{-\pi^2}{\sin^2(\pi z)}.$$

Next, recall that the cotangent function, $\cot(t) = \frac{\cos(t)}{\sin(t)}$, has the derivative:

$$\cot'(t) = -\frac{1}{\sin^2(t)}.$$

Therefore, we have:

$$f(z) = \pi \cot(\pi z) + C.$$

Since both sides are odd functions of z ($f(-z) = -f(z)$), we must have $C = 0$. Thus, we obtain:

$$\pi \cot(\pi z) = \frac{1}{z} + \sum_{n \neq 0} \left(\frac{1}{z-n} + \frac{1}{n} \right).$$

Remark 10.14. *There is an alternative way to achieve convergence in this case, rather than using the general method of polynomial counter-terms. We can combine the terms for $\pm n$:*

$$\frac{1}{z-n} + \frac{1}{z+n} = \frac{2z}{z^2 - n^2}.$$

This results in a convergent series (while the terms $\frac{1}{n} - \frac{1}{n}$ cancel out). Hence, we can rewrite the sum as:

$$\pi \cot(\pi z) = \frac{1}{z} + \sum_{n \geq 1} \frac{2z}{z^2 - n^2}.$$

10.3 Infinite Product Expansions

After studying infinite sum formulas in the spirit of partial fractions, we now turn our attention to infinite products. The convention/definition we adopt for infinite products is as follows:

Definition 10.15. *Let $\prod_{i=1}^{\infty} p_i$ be an infinite product. It converges if:*

1. *At most finitely many terms p_i are zero, and*
2. *The product of the nonzero terms $\prod_{1 \le i \le n, p_i \neq 0} p_i$ converges to a nonzero limit as $n \to \infty$.*

This definition may feel somewhat awkward and less natural than the obvious alternative (where $\prod_{i=1}^{n} p_i$ converges to a limit, which may be zero), but it is more suitable for expressing analytic functions as infinite products.

The requirements ensure the following:

- Adding or removing finitely many factors does not affect the convergence of the product.
- When a convergent product of analytic functions vanishes, it does so to a finite order (i.e., the sum of the orders of the factors that equal zero), and we can factor out the zeroes. (Note that a convergent product of nonzero factors is nonzero by definition!)
- For nonzero products, the convergence of $\prod p_i$ is equivalent to the convergence of $\sum \log p_i$.

Since convergence forces $\log(p_i) \to 0$, i.e., $p_i \to 1$, it is customary to write infinite products in the form $\prod_{n=1}^{\infty}(1 + a_n)$. Convergence of the product is equivalent to the convergence of $\sum \log(1 + a_n)$ (with $a_n \to 0$). We select the principal branch of the logarithm such that $|\mathrm{Im}(\log)| < \pi$.

Moreover, $\sum \log(1 + a_n)$ converges absolutely if and only if $\sum a_n$ converges absolutely. This can be shown using a comparison argument, as either condition implies $a_n \to 0$. For sufficiently large n, we have the inequality $\frac{|a_n|}{2} \le |\log(1 + a_n)| \le 2|a_n|$. When this occurs, we say the product converges absolutely. However, non-absolute convergence may involve more subtle cancellations, and cannot be reduced to the convergence of $\sum a_n$.

The goal is to express a given entire analytic function $f(z)$ as a product that reveals the zeroes of f, just as we write a polynomial in the form $c\prod(z - b_i)^{m_i}$. Since an infinite product of $(z - b_i)$'s does not converge, we aim for a product of factors of the form $\prod_{i=1}^{\infty}\left(1 - \frac{z}{b_i}\right)^{m_i}$ (for $b_i \neq 0$). If f has a zero at $z = 0$, we include the factor z^{m_0}.

If the infinite product converges for all z, and if the convergence is uniform on compact subsets of $\mathbb{C} - \{b_i\}$ (which, by definition, means that $\sum m_i \log\left(1 - \frac{z}{b_i}\right)$ converges uniformly), then it defines an analytic function with the same zeroes

as f. Consequently, the ratio of $f(z)$ and this function is an entire function with no zeroes, and can therefore be written as $e^{g(z)}$ for some entire analytic function $g(z)$.

In summary, our goal is to express $f(z)$ as:

$$f(z) = z^{m_0} e^{g(z)} \prod_{i=1}^{\infty} \left(1 - \frac{z}{b_i}\right)^{m_i}.$$

As in the case of sums, the following questions arise:

- Can we represent given functions in this way?
- When do these expressions converge?
- Given a set $\{b_i\} \subset \mathbb{C}$ with no limit points (i.e., $b_i \to \infty$), can we find an entire function with zeroes of prescribed orders at b_i?

The answers to these questions are analogous to the case of partial fractions. To begin, we look at an example: the function $\sin(\pi z)$.

Example 10.16. *Since* $\sin(\pi z)$ *has zeroes exactly at the integers, a naive guess for its infinite product representation is:*

$$z \prod_{n \neq 0} \left(1 - \frac{z}{n}\right).$$

Unfortunately, the series $\sum \log\left(1 - \frac{z}{n}\right)$ *diverges (just as* $\sum \frac{1}{n}$ *does). To handle this, we cancel the divergence by subtracting the beginning of the Taylor series for each term.*

Here, we have:

$$\log\left(1 - \frac{z}{n}\right) = -\frac{z}{n} - \frac{z^2}{2n^2} - \cdots,$$

so we can consider:

$$\sum \left(\left(1 - \frac{z}{n}\right) + \frac{z}{n}\right),$$

which converges (since $\sum \frac{z^2}{n^2}$ *converges). This gives the product:*

$$z \prod_{n \neq 0} \left(\left(1 - \frac{z}{n}\right) e^{\frac{z}{n}}\right),$$

which converges (by the convergence of $\sum \log(\cdots)$*). Thus, we can write:*

$$\sin(\pi z) = z e^{g(z)} \prod_{n \neq 0} \left(\left(1 - \frac{z}{n}\right) e^{\frac{z}{n}}\right)$$

for some analytic function $g(z)$.

How do we find $g(z)$? The answer is to compare the logarithmic derivatives of both sides. For $z \in \mathbb{C} - \mathbb{Z}$, the logarithmic derivative of a product is the sum of the logarithmic derivatives of the individual factors. Thus, we have:

$$\begin{aligned}
\sin(\pi z) &\longrightarrow \frac{\pi \cos(\pi z)}{\sin(\pi z)} = \pi \cot(\pi z), \\
z &\longrightarrow \frac{1}{z}, \\
\prod_{n \neq 0} \left(\left(1 - \frac{z}{n}\right) e^{\frac{z}{n}} \right) &\longrightarrow \sum_{n \neq 0} \left(\frac{-1/n}{1 - z/n} + \frac{1}{n} \right) = \sum_{n \neq 0} \left(\frac{1}{z - n} + \frac{1}{n} \right), \\
e^{g(z)} &\longrightarrow g'(z).
\end{aligned}$$

Thus, we obtain:

$$\pi \cot(\pi z) = \frac{1}{z} + g'(z) + \sum_{n \neq 0} \left(\frac{1}{z - n} + \frac{1}{n} \right).$$

Using the previously derived formulas, we find that $g'(z) = 0$, so $e^{g(z)}$ is a constant. To determine this constant c, we divide both sides by z and evaluate at $z = 0$:

$$\lim_{z \to 0} \frac{\sin(\pi z)}{z} = c.$$

Thus, $c = \pi$, and we conclude:

$$\sin(\pi z) = \pi z \prod_{n \neq 0} \left(\left(1 - \frac{z}{n}\right) e^{\frac{z}{n}} \right).$$

By grouping terms corresponding to $+n$ and $-n$, we can also write:

$$\sin(\pi z) = \pi z \prod_{n \geq 1} \left(1 - \frac{z^2}{n^2} \right).$$

Remark 10.17. *Earlier and now, the series $\sum_{n \neq 0} \frac{1}{z-n}$ and $\sum_{n \neq 0} \log\left(1 - \frac{z}{n}\right)$ are considered divergent because we must think of $\sum_{n \neq 0} = \sum_{n > 0} + \sum_{n < 0}$, and both series are divergent. The simpler rewriting by grouping $\pm n$ together amounts to the observation that for these specific divergent series, there is a convergent rearrangement:*

$$\lim_{N \to \infty} \left(\sum_{n=-N, n \neq 0}^{N} a_n \right) = \lim_{N \to \infty} \left(\sum_{n=1}^{N} (a_n + a_{-n}) \right).$$

This series $a_1 + a_{-1} + a_2 + a_{-2} + \cdots$ converges non-absolutely. However, rearranging non-absolutely convergent series is not a benign operation; it can change the value of the sum. In fact, for series of real numbers, you can make the sum take any value you wish! (See Rudin's Theorem 3.54)

Theorem 10.18 (The General Existence Theorem)**.** *Given a subset* $\{b_1, b_2, \dots\} \subset \mathbb{C}$ *with* $|b_j| \to \infty$ *(i.e., no limit points) and multiplicities* $m_j \geq 1$*, there exists an entire analytic function* $f(z)$ *with zeroes exactly at the points* b_j*, with order* m_j *at each.*

The proof follows the same steps as for partial fractions: we want to modify the sum $\sum m_j \log\left(1 - \frac{z}{b_j}\right)$ to achieve convergence. As before, we do this by subtracting part of the Taylor series expansion $(\star)$:

$$\log\left(1 - \frac{z}{b_j}\right) = -\frac{z}{b_j} - \frac{z^2}{2b_j^2} - \cdots,$$

and stopping at some degree d_j. We then consider the infinite product:

$$z^{m_0} \prod_j \left[\left(1 - \frac{z}{b_j}\right) e^{\frac{z}{b_j} + \frac{1}{2}\left(\frac{z}{b_j}\right)^2 + \cdots + \frac{1}{d_j}\left(\frac{z}{b_j}\right)^{d_j}}\right]^{m_j}.$$

As with partial fractions, the appropriate choice of d_j's ensures that the remainders in $(\star)$ form a series such that $\sum m_j r_j(z)$ converges uniformly on compact subsets. Thus, the infinite product converges (uniformly).

Corollary 10.19. *Every meromorphic function on* $\mathbb{C}$ *is the quotient of two entire analytic functions.*

Proof. Suppose f has poles at $\{b_j\}$ with orders m_j. By the General Existence Theorem, there exists an entire function $g(z)$ with zeroes precisely at b_j, with order m_j at each. Thus, the function $h(z) = g(z)f(z)$ is analytic everywhere (the zeroes of g cancel the poles of f), and we can write $f(z) = \frac{h(z)}{g(z)}$. □

10.4 Gamma and Zeta Functions

This section explores an application of infinite sums and products: constructing new functions.

Warm-up: the partition generating function. Let $p(n)$ denote the number of partitions of n, i.e., the number of ways to express n as an unordered sum of positive integers (with the convention $p(0) = 1$).

$$\begin{aligned} 1, &\quad p(1) = 1 \\ 2 = 1 + 1, &\quad p(2) = 2 \\ 3 = 2 + 1 = 1 + 1 + 1, &\quad p(3) = 3 \\ 4 = 3 + 1 = 2 + 2 = 2 + 1 + 1 = 1 + 1 + 1 + 1, &\quad p(4) = 5 \end{aligned}$$

and so on.

This function has many remarkable properties, such as those related to arithmetic (e.g., Ramanujan's result: $p(5k+4) \equiv 0 \pmod 5$). However, our main goal here is to examine the growth rate of $p(n)$: Is it polynomial or exponential?

One way to approach this is to introduce the generating function

$$P(z) = \sum_{n=0}^{\infty} p(n) z^n$$

and investigate its properties (such as the radius of convergence). The key formula for this function is its product expansion (Euler, 1753):

$$P(z) = \sum_{n=0}^{\infty} p(n) z^n = \prod_{n=1}^{\infty} \frac{1}{1-z^n}.$$

To understand this, we express the product as the infinite series:

$$(1 + z + z^2 + \dots)(1 + z^2 + z^4 + \dots)(1 + z^3 + z^6 + \dots)\dots$$

A partition of n as a sum of a_1 ones, a_2 twos, etc., corresponds to the contribution to the coefficient of z^n from multiplying z^{a_1} in the first factor, z^{2a_2} in the second, and so on. Therefore, the total coefficient of z^n is indeed $p(n)$.

This infinite product expansion, and the comparison between $\sum \log(1-z^n)$ and $\sum z^n$, shows that $P(z)$ is well-defined and analytic in the unit disk $D = \{z \mid |z| < 1\}$. However, we also observe that since the factors have poles at all roots of unity (which form a dense subset of the unit circle, $e^{2\pi i\alpha}$, where $\alpha \in \mathbb{Q}$), there is no way to extend $P(z)$ beyond the unit disk. This tells us that the radius of convergence is 1, but a much more detailed analysis of $P(z)$ provides additional information: specifically, $p(n) \sim \frac{1}{4n\sqrt{3}} \exp\left(\pi\sqrt{\frac{2n}{3}}\right)$ (Hardy-Ramanujan, 1918).

Next, let us discuss the Gamma function.

Question 1: Does there exist a meromorphic function that generalizes $n!$ to non-negative integers?

Since $n! = n \times (n-1)!$, the functional identity we would hope for is $F(z) = zF(z-1)$. This cannot be a polynomial, however, because comparing the zeros on both sides of the identity reveals that the zeros of $F(z)$ are those of $F(z-1)$ (i.e., the zeros of F shifted by 1) plus one additional zero at $z = 0$. This implies that if F is an entire function, it must have zeros at all non-negative integers, which is inconsistent with the goal of generalizing $n!$. A better approach is to seek a meromorphic function with poles at the negative integers (and no zeros).

Question 2: Is there an entire function whose zeros are exactly the negative integers?

Yes, we can construct such a function:

$$G(z) = \prod_{n=1}^{\infty} \left(\left(1 + \frac{z}{n}\right) e^{-z/n} \right)$$

This ensures convergence of the series $\sum_{n\geq 1} \left(\log\left(1+\frac{z}{n}\right) - \frac{z}{n}\right)$. Note that $zG(z)G(-z) = \frac{1}{\pi}\sin(\pi z)$, as previously established.

What functional equation does G satisfy? We observe that $G(z-1)$ has zeros at $z = 0, -1, -2, \ldots$, which are the same as the zeros of $zG(z)$. Hence, the function $\frac{G(z-1)}{zG(z)}$ is entire and has no poles at its zeros, implying that it must be of the form $e^{\gamma(z)}$ for some entire function $\gamma(z)$.

Thus, we have the equation

$$G(z-1) = zG(z)e^{\gamma(z)}.$$

To find $\gamma(z)$, we take the logarithmic derivative of both sides:

$$\frac{G'(z)}{G(z)} = \frac{1}{z} + \frac{G'(z)}{G(z)} + \gamma'(z) \implies \gamma'(z) = 0 \implies \gamma(z) = \text{constant}.$$

This constant is known as **Euler's constant**:

$$G(0) = 1 = G(1)e^{\gamma} \implies \gamma = -\log G(1) = \sum_{n=1}^{\infty} \left(\frac{1}{n} - \log\left(\frac{n+1}{n}\right)\right) = \lim_{n\to\infty} \left(1 + \frac{1}{2} + \cdots + \frac{1}{n} - \log(n)\right) \approx 0.57$$

To eliminate the factor e^{γ}, we define a new function $H(z) = e^{\gamma z}G(z)$. This yields the equation:

$$H(z-1) = e^{\gamma z}e^{-\gamma}G(z-1) = e^{\gamma z}zG(z) = zH(z).$$

Thus, we have

$$H(z) = e^{\gamma z} \prod_{n=1}^{\infty} \left(\left(1 + \frac{z}{n}\right) e^{-z/n} \right) = \prod_{n=1}^{\infty} \left(1 + \frac{z}{n}\right) \left(1 + \frac{1}{n}\right)^{-z}.$$

Finally, we define the Gamma function.

Definition 10.20. *The **Gamma function** is defined as*

$$\Gamma(z) = \frac{1}{zH(z)} := \frac{1}{H(z-1)}.$$

Here are some properties of the Gamma function:

Proposition 10.21.
- $\Gamma(z)$ *is a meromorphic function with simple poles at* $z = 0, -1, -2, \ldots$ *and no zeros.*
- *The functional form of* $\Gamma(z)$ *is given by*

$$\Gamma(z) = \frac{e^{-\gamma z}}{z} \prod_{n=1}^{\infty} \left(1 + \frac{z}{n}\right)^{-1} e^{z/n} = \frac{1}{z} \prod_{n=1}^{\infty} \left(1 + \frac{1}{n}\right)^{z} \left(1 + \frac{z}{n}\right)^{-1}.$$

- *The functional equation* $\Gamma(z+1) = z\Gamma(z)$ *holds.*
- *Since* $\Gamma(1) = 1$*, we have* $\Gamma(n) = (n-1)!$ *for all* $n \in \mathbb{Z}_{>0}$*.*
- *From the identity* $\pi z G(z) G(-z) = \sin(\pi z)$*, we derive the reflection formula*

$$\Gamma(z)\Gamma(1-z) = \frac{\pi}{\sin(\pi z)}.$$

Theorem 10.22 (Stirling's Formula)**.**

$$\Gamma(z) \sim \sqrt{2\pi} z^{z-\frac{1}{2}} e^{-z} \quad \textit{as} \quad \textit{Re}(z) \to \infty.$$

Remark 10.23. *We will skip the proof as it is quite involved. For a detailed explanation, refer to Ahlfors 5.2.5.*

This approximation implies that $n! \sim \sqrt{2\pi n} n^n e^{-(n+1)}$*, as seen in Homework 7.*

Next, consider the integral representation of $\Gamma(z)$:

$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t}\, dt \quad \text{for} \quad \text{Re}(z) > 0.$$

Integration by parts shows that this integral satisfies the same functional identity as $\Gamma(z)$. The ratio of the two is 1-periodic and entire, and Stirling's formula implies that this function is bounded, hence constant (equal to 1 when evaluated at positive integers).

Many other fascinating formulas exist for the Gamma function, such as Legendre's duplication formula:

$$\sqrt{\pi}\Gamma(2z) = 2^{2z-1}\Gamma(z)\Gamma\left(z + \frac{1}{2}\right).$$

(See Ahlfors for more details.)

Let us now turn our discussion to the Riemann zeta function. We have seen how to encode a sequence of numbers a_n into a generating function, specifically

a power series $\sum a_n z^n$. However, one can also try a different approach: the Dirichlet series, given by

$$f(s) = \sum_{n=1}^{\infty} \frac{a_n}{n^s}$$

(where, for traditional reasons, the variable is denoted by s rather than z).

The simplest such series is the Riemann zeta function defined as

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}.$$

This series converges absolutely for $\mathrm{Re}(s) > 1$ and uniformly on the set $\{\mathrm{Re}(s) \geq 1 + \epsilon\}$ for any $\epsilon > 0$. Consequently, $\zeta(s)$ is an analytic function on $\{\mathrm{Re}(s) > 1\}$.

Proposition 10.24. *Although the series does not converge for Re$(s) < 1$, the function $\zeta(s)$ can be extended to a meromorphic function on the entire complex plane, with a pole at $s = 1$.*

The primary questions about $\zeta(s)$ concern its behavior in regions of the complex plane where the series diverges. For example, its number-theoretic significance is given by the Euler product:

$$\zeta(s) = \prod_{p \text{ prime}} \frac{1}{1 - p^{-s}},$$

where we use the identity $\frac{1}{1-p^{-s}} = 1 + \frac{1}{p^s} + \frac{1}{p^{2s}} + \cdots$, along with the prime factorization.

Because of this representation, the behavior of $\zeta(s)$ as a complex analytic function reflects the properties of the primes.

Example 10.25. *The fact that $\sum \frac{1}{n}$ diverges is equivalent to the following three statements:*

$$\iff \textit{poles of } \zeta(s) \textit{ at } s = 1 \iff \textit{there are infinitely many primes } p,$$

and the series

$$\sum \log\left(\frac{1}{1 - p^{-1}}\right) \sim \sum \frac{1}{p}$$

diverges.

However, there are much deeper facts. The location of the zeros of $\zeta(s)$ implies estimates on the error term in the classical approximation for the prime counting function:

$$\pi(x) \sim \#\{\text{primes } p \leq x\} \sim \frac{x}{\log x} + O\left(\frac{x}{\log^2 x}\right),$$

which is the Prime Number Theorem. This is the subject of the Riemann Hypothesis.

Returning to complex analysis: the function $\zeta(s)$ is intimately related to the Gamma function $\Gamma(s)$, since

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt,$$

and

$$n^{-s}\Gamma(s) = \int_0^\infty t^{s-1} e^{-nt}\, dt.$$

This follows by a change of variables nt, where $t^{z-1}dt \to n^z t^{z-1} dt$.

Summing over $n \geq 1$, we obtain for $\mathrm{Re}(s) > 1$:

$$\zeta(s)\Gamma(s) = \int_0^\infty \frac{t^{s-1}}{e^t - 1}\, dt,$$

since

$$\sum_{n=1}^{\infty} e^{-nt} = \frac{e^{-t}}{1 - e^{-t}} = \frac{1}{e^t - 1}.$$

This allows us to re-express $\zeta(s)$ as a path integral: $\frac{(-z)^{s-1}}{e^z - 1}$ has branching behavior at $z = 0$ and poles at $2\pi i n, n \in \mathbb{Z}$. So, we have:

$$\int_C \frac{(-z)^{s-1}}{e^z - 1} = -\int_0^\infty \frac{x^{s-1} e^{-i\pi(s-1)}}{e^x - 1}\, dx + \int_0^\infty \frac{x^{s-1} e^{i\pi(s-1)}}{e^x - 1}\, dx,$$

which simplifies to:

$$= 2i \sin(\pi(s-1))\zeta(s)\Gamma(s).$$

(Cauchy's theorem implies the integral is independent of $\epsilon \in (0, 2\pi)$ and for $\mathrm{Re}(s) > 1$, we have $\lim_{\epsilon > 0} \int_{S^1(\epsilon)} = 0$.)

Since $\Gamma(s)\Gamma(1-s) = \frac{\pi}{\sin(\pi s)}$, we get:

$$\zeta(s) = -\frac{\Gamma(1-s)}{2i\pi} \int_C \frac{(-z)^{s-1}}{e^z - 1}\, dz \qquad (\star)$$

The point is: the right-hand side is defined and meromorphic for all $s \in \mathbb{C}$! The integral converges at infinity because e^z in the denominator grows much faster than $|z|^{s-1}$. Analytic dependence on s follows from our usual tricks for integral formulas (such as differentiating under the integral).

Since $\Gamma(1-s)$ has poles at $1 - s \in \{0, -1, -2, \dots\}$, i.e., at $s = \{1, 2, 3, \dots\}$, the only possible poles of $\zeta(s)$ are at $s = 1, 2, 3, \dots$. However, for $s \geq 2$, the series $\sum \frac{1}{n^s}$ converges, so the pole of $\Gamma(1-s)$ is canceled by the vanishing of the integral (no branching behavior for $s \in \mathbb{Z}$).

Corollary 10.26. *$\zeta(s)$ extends to an entire meromorphic function, whose only pole is a simple pole at $s = 1$.*

Further consideration of the integral formula ($\star$) yields the "functional equation" for $\zeta(s)$:

Theorem 10.27.

$$\zeta(s) = 2^s \pi^{s-1} \Gamma(1-s)\zeta(1-s).$$

Remark 10.28. *This is proved by further manipulation of the integral* ($\star$)*, and by closing the path in* $\mathbb{C}$*, see Ahlfors 5.4.3.*

This is important: we know that $\zeta(s)$ has no zeros in the half-plane $\mathrm{Re}(s) > 1$ (as seen from the product expansion $\zeta(s) = \prod_p \frac{1}{1-p^{-s}}$, which converges for $\mathrm{Re}(s) > 1$), so this equation determines the behavior of ζ in the half-plane $\mathrm{Re}(s) < 0$. Specifically, it has simple zeros at $s = -2, -4, -6, \ldots$ and no other zeros.

The remaining zeros lie in the "critical strip" $0 < \mathrm{Re}(s) < 1$; the **Riemann Hypothesis** states that these zeros all lie on the line $\mathrm{Re}(s) = \frac{1}{2}$. This has been verified experimentally for the first few million zeros (starting with $\frac{1}{2} \pm 14.134725\cdots i$, $\frac{1}{2} \pm 21.022039\cdots i$, etc.) and is widely believed to be true (which has implications for the distribution of prime numbers), but a proof remains elusive. (The Clay Mathematics Institute offers a $1 million prize for a proof or disproof.)

10.5 Abelian Integrals and Elliptic Functions

Riemann surfaces were historically introduced to handle the multivalued nature of certain algebraic functions and their integrals. For example, consider the integral

$$I = \int_{z_0}^{z_1} \frac{dz}{\sqrt{z^2+1}}.$$

One might evaluate this integral using trigonometric substitutions, such as $z = \sinh(w)$, but a more elegant approach is to interpret it as a **path integral on a Riemann surface**. This is because the function $\sqrt{z^2+1}$ is multivalued: there are two possible values whenever $z \notin \{\pm i\}$. Its graph forms a three-sheeted covering space over the complex plane excluding the points $\pm i$, with $w = \pm\sqrt{z^2-1}$. If we vary z along a path, say a circle around one of $\pm i$, the lift of this path to the covering space changes sheets. Starting at a point w, the path eventually returns to $-w$.

Thus, we introduce the set

$$\Sigma = \{(z, w) \in \mathbb{C}^2 \mid w^2 = z^2 + 1\},$$

and now view z and w as single-valued analytic functions on Σ, rather than as multivalued functions on $\mathbb{C}$. The set Σ is an example of a complex manifold. Around each point of Σ, we can use either w or z as a local coordinate and

express all functions as analytic functions of it. In particular, the integral is now best understood as

$$\int_{p_0}^{p_1} \frac{dz}{w},$$

where $p_0 = (z_0, w_0)$ and $p_1 = (z_1, w_1)$ are points on Σ.

While this may seem like an unnecessary complication if you already have a clear idea of how to evaluate the integral, it can often provide considerable insight into the problem.

The remarkable fact here is that Σ is biholomorphic to a domain in the complex plane. Explicitly, in terms of the Riemann sphere $S = \mathbb{C} \cup \{\infty\}$, we have the following inverse analytic bijections:

$$S - \{\pm 1\} \xrightarrow{\cong} \Sigma = \{(z, w) \mid w^2 = z^2 + 1\}$$

$$\lambda \mapsto \left(\frac{2\lambda}{1 - \lambda^2}, \frac{1 + \lambda^2}{1 - \lambda^2} \right)$$

$$(z, w) \mapsto \frac{w - 1}{z}.$$

Therefore, we can transform our path integral on Σ into one on $S - \{\pm 1\}$ by making the change of variables

$$w = \frac{1 + \lambda^2}{1 - \lambda^2}, \quad dz = \frac{2(1 + \lambda^2)}{(1 - \lambda^2)^2}\, d\lambda.$$

Thus, the integral

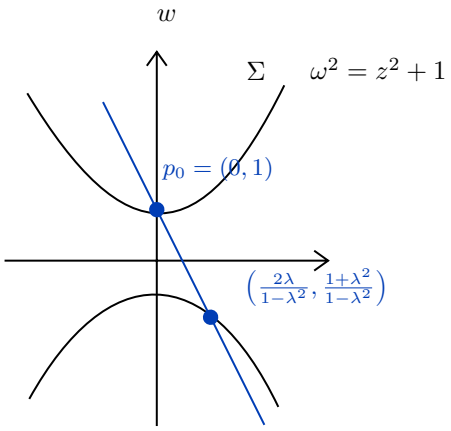
$$\int_{p_0}^{p_1} \frac{dz}{w}$$

becomes

$$\int_{\lambda_0}^{\lambda_1} \frac{2\, d\lambda}{1 - \lambda^2},$$

which is easier to handle using partial fractions.

What is the geometric interpretation of this change of variables? The set $\Sigma = \{(z, w) \in \mathbb{C}^2 \mid w^2 = z^2 + 1\}$ is defined by an algebraic equation of degree 2. The intersection of Σ with a complex line in $\mathbb{C}^2$ typically consists of two points. Therefore, we can project Σ along the family of lines through a fixed point $p_0 \in \Sigma$ (e.g., $(0, 1)$). Each of these lines intersects Σ at the point p_0 and typically one other point. This idea is conceptually similar to stereographic projection of the sphere $S^2 \subset \mathbb{R}^3$, where degree-2 equations come into play. However, in this case, we are working in $\mathbb{C}^2$ rather than $\mathbb{R}^3$.



The line with slope λ through p_0 has the equation $w = \lambda z + 1$. Substituting this into the equation $w^2 = z^2 + 1$ results in a degree-2 equation in z (with coefficients depending on λ), which always has $z = 0$ as one of its roots. This makes it especially easy to find the other root.

$$\begin{aligned} &(\lambda z + 1)^2 = z^2 + 1 \\ \rightsquigarrow &(\lambda^2 - 1)z^2 + 2\lambda z = 0 \\ \rightsquigarrow &z = 0 \text{ or } z = \frac{2\lambda}{1 - \lambda^2} \end{aligned}$$

Additionally, every point $p \in \Sigma$ ($p \neq p_0$) arises from this construction by taking the line $(p_0 p)$. Special cases include:

- For $\lambda = 0$, the line L_λ is tangent to Σ at p_0, resulting in a double root $z = 0$.
- For $\lambda = \pm 1$, the other intersection of L_λ and Σ disappears ("at ∞").
- To obtain the point $(0, -1) \in \Sigma$, we must allow $\lambda = \infty$.

This construction provides a biholomorphism

$$S - \{\text{finite set}\} \xrightarrow{\sim} \Sigma$$

given by rational functions (assuming Σ is a rational curve; the term "curve" refers to it being complex 1-dimensional), even though Σ appears as a surface in real 2-dimensions.

This process allows us to evaluate path integrals on algebraic curves $\Sigma \subset \mathbb{C}^2$ defined by any quadratic polynomial $Q(z, w) = 0$; however, complications arise when we try to extend this method.

Question: Calculate the arc length of a portion of the ellipse $x^2 + \frac{y^2}{2} = 1$ between (x_0, y_0) and (x_1, y_1). If we write $y = \pm\sqrt{2(1-x^2)}$ and use

$$\int_{x_0}^{x_1} \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx,$$

we arrive at an integral of the form

$$\int_{x_0}^{x_1} \sqrt{\frac{1+x^2}{1-x^2}}\, dx.$$

Alternatively, using parametric length, we obtain

$$\int_{\theta_0}^{\theta_1} \sqrt{1 + \cos^2(\theta)}\, d\theta.$$

By manipulating this further, we can reduce it to an expression like

$$\int \frac{dx}{\sqrt{1-x^4}}.$$

However, none of these "elliptic integrals" can be expressed in terms of known functions. Early 19th-century mathematicians were at an impasse until Riemann, Abel, and others provided the right perspective. Riemann surfaces are necessary to make sense of these integrals. (This is a topic at the intersection of complex analysis, topology, and algebraic geometry!) Thus, we now consider the graph of $\sqrt{1-z^4}$, given by the equation

$$\Sigma = \{(z, w) \in \mathbb{C}^2 \mid w^2 = z^4 - 1\}.$$

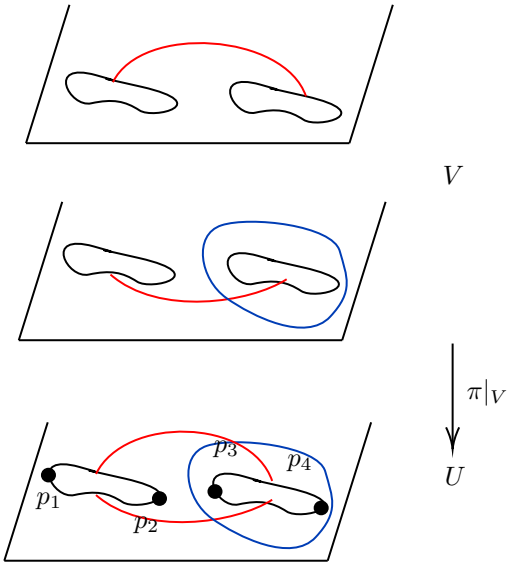
Claim: The reason this case differs from the previous one is that Σ is not an open subset of the Riemann sphere, but rather an open subset of a torus (an "elliptic curve"—the term comes from the problem of elliptic integrals and the associated challenges).

To understand this, we project onto the z-coordinate: $(z, w) \mapsto z$. This map is a "branched covering"—a two-sheeted covering map—after we remove the roots p_i of the polynomial $z^4 - 1$ (which are ± 1 and $\pm i$) from $\mathbb{C}$, and the corresponding points $q_i = (p_i, 0)$ from Σ. Hence, the map

$$\Sigma - \{q_i\} \xrightarrow{\pi} \mathbb{C} - \{p_i\}, \quad (z, w) \mapsto z,$$

is a **2:1 covering**.

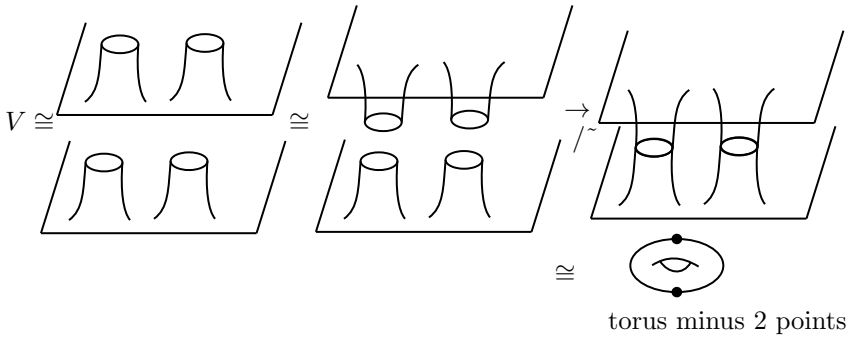
The points p_i are the **branch points**: the lift of a small circle around p_i is a path that ends up on the opposite sheet of where it started (i.e., $w \mapsto -w$). In general, a loop in $\mathbb{C} - \{p_i\}$ lifts to a loop in $\Sigma - \{q_i\}$ if and only if the sum of its winding numbers around p_1, p_2, p_3, p_4 is even.



Consider two arcs γ and γ' in $\mathbb{C}$, connecting p_1 to p_2 and p_3 to p_4 (for example), and let $U = \mathbb{C} - (\gamma \cup \gamma')$. Then, any loop in U has an even total winding number, so it lifts to a loop in Σ.

Hence, the restricted covering map from $V = \pi^{-1}(U)$ to U is trivial. Specifically, V decomposes as $V = V_+ \sqcup V_-$, and the map $\pi_{|V_\pm} : V_\pm \xrightarrow{\sim} U$ is a bijection. This makes the slits in these planes more visible: by adding back the missing arcs $\gamma \cup \gamma'$, the lift of a path in $\mathbb{C}$ jumps between the two sheets $V_\pm$ each time it crosses $\gamma \cup \gamma'$.

Thus, Σ is obtained from V by attaching one side of each slit in each sheet to the other side of the same slit in the other plane.



It is possible to compactify $\mathbb{C}$ to the sphere S and Σ to a torus $\overline{\Sigma}$ by adding two preimages of ∞.

The implication for complex analysis is that, since $\overline{\Sigma}$ is not simply connected, path integrals on it depend on the path of integration.

Returning to our integral

$$\int \frac{dz}{w}$$

on $\Sigma = \{w^2 = z^4 - 1\}$ (or another polynomial of degree 3 or 4 with simple roots), we make the following observations:

- The expression $\frac{dz}{w}$ is an analytic 1-form on $\overline{\Sigma}$, with no poles or zeroes. Specifically, at $(z, w) = (p_i, 0)$, the local coordinate on Σ is actually w, not z. Since $w^2 = P(z)$, we have $2w\,dw = P'(z)\,dz$, which implies

  $$\frac{dz}{w} = \frac{2\,dw}{P'(z)},$$

  and thus there is no pole.

- The integral $\int_{p_0}^{p_1} \frac{dz}{w}$ is invariant under path homotopy (by the Cauchy theorem), but it depends on the homotopy class. If we choose loops α_1, α_2 that generate $\pi_1\left(\overline{\Sigma}\right) \simeq \mathbb{Z}^2$, a change in the homotopy class modifies the value of the integral by an integer linear combination of the periods $w_1 = \int_{\alpha_1} \frac{dz}{w}$ and $w_2 = \int_{\alpha_2} \frac{dz}{w}$. Given two paths γ and γ' from p_0 to p_1, we have

  $$[\gamma - \gamma'] = m_1[\alpha_1] + m_2[\alpha_2]$$

  for some $m_1, m_2 \in \mathbb{Z}$, which implies

  $$\int_\gamma - \int_{\gamma'} = m_1 w_1 + m_2 w_2.$$

- The function

  $$\int_{p_0}^{p} \frac{dz}{w} = F(p)$$

  defines an analytic mapping

  $$F : \overline{\Sigma} \to \mathbb{C}/\mathbb{Z}w_1 \oplus \mathbb{Z}w_2,$$

  which has the following properties:

  - It cannot be expressed in terms of elementary functions.
  - It has everywhere a nonzero derivative, so F is a local homeomorphism, and in fact, a covering map.
  - By winding number arguments (for complex analysts) or by studying the map on fundamental groups (for topologists), we can show that $|F^{-1}(c)| = 1$ for all c, meaning that F is a biholomorphism.

Problem 10.29. *What is the inverse of F?*

Solution. The inverse of F is a doubly periodic function, which is approximately the Weierstrass $\wp$-function. □

10.6 The Weierstrass $\wp$-function

Consider double-periodic functions $f(z+w_1) = f(z+w_2) = f(z)$. If f is analytic, then it must be bounded and hence constant. Therefore, the only interesting such functions are meromorphic. The residue formula for integrating around a large parallelogram implies that the sum of residues in the fundamental domain must be zero (since the path integral is linear in N, while the sum of residues is quadratic in N). This leads to the conclusion that we cannot have just a single pole of order 1 in the fundamental domain.

The simplest such functions either have one pole of order 2 or two poles of order 1 in the parallelogram defined by w_1 and w_2. Weierstrass' starting point has a pole of order 2, with vanishing residue. Up to translation, we can place the pole at $z = 0$ with the polar part $\frac{1}{z^2}$. Following our study of infinite sums and how to achieve convergence, we obtain the Weierstrass $\wp$-function:

$$\wp(z) = \frac{1}{z^2} + \sum_{w \neq 0} \left(\frac{1}{(z-w)^2} - \frac{1}{w^2} \right)$$

where $w = n_1 w_1 + n_2 w_2$ and $(n_1, n_2) \in \mathbb{Z}^2 - \{(0,0)\}$.

This series converges uniformly on compact sets, since the series $\sum_{w \neq 0} \frac{1}{|w|^3}$ converges. The derivative $\wp'(z) = -2 \sum_w \frac{1}{(z-w)^3}$ is obviously periodic, so $P(z+w_1) - \wp(z)$ and $\wp(z+w_2) - \wp(z)$ are both constant. Since $\wp(z)$ is an even function ($\wp(-z) = \wp(z)$), we can evaluate it at $z = \frac{w_1}{2}$ and $z = \frac{w_2}{2}$ to conclude that $\wp(z)$ is periodic.

Next, working on the Laurent expansions at $z = 0$, we find:

$$\wp(z) = \frac{1}{z^2} + \frac{g_2}{20} z^2 + \frac{g_3}{28} z^4 + \dots$$

for some constants $g_2, g_3 \in \mathbb{C}$ (depending on w_1, w_2). Notice that the constant term vanishes, and the odd terms vanish because $\wp$ is even. Thus, we have:

$$\begin{aligned} \wp'(z) &= \frac{-2}{z^3} + \frac{g_2}{10} z + \frac{g_3}{7} z^3 + \dots \\ \implies \wp'(z)^2 &= \frac{4}{z^6} + \frac{3g_2}{5z^2} + \frac{3g_3}{7} + \dots \\ \implies \wp'(z)^2 &= 4\wp(z)^3 - g_2 \wp(z) - g_3. \end{aligned}$$

The polar parts match, so we equate the entire expressions up to a constant, and the constant terms match as well. The outcome is that the map $z \mapsto (\wp(z), \wp'(z))$ gives a biholomorphism:

$$\mathbb{C}/\mathbb{Z}w_1 + \mathbb{Z}w_2 \xrightarrow{\sim} \{(x,y) \in \mathbb{C}^2 \mid y^2 = 4x^3 - g_2 x - g_3\} \cup \{\infty\},$$

which is another elliptic curve.

Additionally, we have $d\wp(z) = \wp'(z)\,dz$, which gives $dz = \frac{d\wp(z)}{\wp'(z)} = \frac{dx}{y}$. Thus, the inverse function is given by:

$$\int \frac{dx}{y} = \int \frac{dx}{\sqrt{4x^3 - g_2 x - g_3}}.$$

This is almost the same as the previous example, except this one has one of the four branch points at ∞, unlike our previous example where all four poles $p_i \in \mathbb{C}$. Simple coordinate transformations by rational functions allow us to switch between the two cases.

Finally, consider a polynomial $f(x, y) \in \mathbb{Q}[x, y]$ with rational coefficients.

Problem 10.30. *How many rational solutions* $\{(x, y) \in \mathbb{Z} \mid f(x, y) = 0\}$ *are there?*

Solution. In fact, the answer is governed by the topology of the Riemann surface $\overline{\Sigma}$ obtained by compactifying $\Sigma = \{(x, y) \in \mathbb{C}^2 \mid f(x, y) = 0\}$, specifically by its genus g. If $g = 0$ (a rational curve, isomorphic to $S = \mathbb{C} \cup \{\infty\}$) or $g = 1$ (an elliptic curve, isomorphic to $\mathbb{C}/\mathbb{Z}w_1 + \mathbb{Z}w_2$), then algebraic operations (e.g., addition in an elliptic curve) yield new rational solutions from known ones. In this case, the number of solutions over $\mathbb{Q}$ can be infinite. □

Theorem 10.31 (Faltings)**.** *If* $g \geq 2$*, then there are only finitely many rational solutions.*

At this point, we have brought together algebra, analysis, topology, geometry, and number theory! This is a good place to end Math 55.