Notes on the Mathematical Tripos

Sky Wilshaw

Part II

University of Cambridge
2020–2024

Contents

I. Algebraic Topology

Lectured in Michaelmas 2022 by Prof. J. Rasmussen

This course is an introduction to the basic ideas of algebraic topology. In the first half of the course, we study an invariant of based topological spaces called the fundamental group. This invariant associates a group to a topological space (with a basepoint). It has the important property that a continuous map between topological spaces induces a homomorphism between their fundamental groups, and that the composition of two maps is mapped to the composition of the corresponding homomorphisms. In slightly fancier language, the fundamental group determines a functor from the category of based topological spaces to the category of groups. The phenomena that the fundamental group detects are essentially one-dimensional; it measures the failure of closed loops in the space to bound two-dimensional disks.

In the second half of the course, we study another functor from spaces to groups, called homology, which enables us to understand higher-dimensional 'holes' in the space. There are many different ways to define homology; we use a relatively concrete one called simplicial homology, which makes sense for a somewhat restricted class of spaces. The notion of homology plays a central role in modern geometry and topology as well as in many branches of algebra and number theory.

Using these invariants we can distinguish various spaces from each other; for example, we can prove that $\mathbb{R}^n$ is not homeomorphic to $\mathbb{R}^m$ when n is not equal to m. We are also able to prove the fundamental theorem of algebra, and to show that certain maps from a space to itself (for example, any continuous map from the closed n-dimensional disk to itself) must have fixed points.

Contents

1. Motivation

1.1. Invariants

Topological spaces are difficult to study on their own, and so we will assign algebraic invariants to these spaces which allow us to reason more easily about these spaces. To a topological space X, a 'numerical invariant' is a number $g(X) \in \mathbb{R} \cup \{\infty\}$ such that $X \simeq Y$ (where $\simeq$ denotes homeomorphism) implies $g(X) = g(Y)$. An example of a numerical invariant is the number of path-connected components of X. An algebraic invariant is a group $G(X)$ assigned to a topological space X such that $X \simeq Y$ implies $G(X) \simeq G(Y)$, where here $\simeq$ denotes isomorphism. We will construct two kinds of such invariants: the fundamental group, and invariants related to homology. The invariants we construct will behave nicely under maps: if $f : X \to Y$ is a continuous map, we induce a homomorphism $f_\star : G(X) \to G(Y)$. We will prove the following model theorems.

- If $\mathbb{R}^n \simeq \mathbb{R}^m$, then $n = m$.
- If $f : D^n \to D^n$ is continuous, then there exists $x \in D^n$ with $f(x) = x$.

The above theorems are easy to prove in the case $n = 1$ by appealing to path-connectedness and the intermediate value theorem. Our study allows us to prove similar things about these higher dimensional cases, among other things.

1.2. Notation

- A *space* is a topological space.
- A *map* is a continuous function, unless defined otherwise.
- If X and Y are spaces, $X \simeq Y$ means that X and Y are homeomorphic.
- If G and H are groups, $G \simeq H$ means that G and H are isomorphic.
- Some common spaces include:
  - The one-point space $\{\bullet\}$;
  - $I = [0, 1] \subset \mathbb{R}$;
  - $I^n = \underbrace{I \times \cdots \times I}_{n \text{ times}}$, the n-dimensional closed unit cube;
  - $D^n = \{v \in \mathbb{R}^n \mid \|v\| \leq 1\}$, the n-dimensional closed unit disk (note that $I^n \simeq D^n$);
  - $S^{n-1} = \{v \in \mathbb{R}^n \mid \|v\| = 1\}$, the $(n-1)$-dimensional unit sphere.
- Common maps include:
  - If X is a space, the identity map $\mathrm{id}_X : X \to X$ is defined by $x \mapsto x$;
  - If X and Y are spaces with $p \in Y$, the constant map $c_{X,p} : X \to Y$ is defined by $x \mapsto p$.

2. Homotopy

2.1. Definition

Definition. Let $f_0, f_1 : X \to Y$ be continuous. We say f_0 is *homotopic to* f_1, written $f_0 \sim f_1$, if there exists a continuous $H : X \times I \to Y$ with $H(x, 0) = f_0(x)$ and $H(x, 1) = f_1(x)$.

We can think of H as a path from f_0 to f_1 in the set $\mathrm{Hom}(X, Y)$ of functions $X \to Y$, which is continuous under a topology that will not be defined here.

Lemma (Gluing lemma)**.** Let $X = C_1 \cup C_2$, where C_1, C_2 are closed in X. Let $f : X \to Y$ be a function (that may be not continuous), such that $f|_{C_1}$ and $f|_{C_2}$ are continuous. Then f is continuous.

Proof. It suffices to show that the preimage of a closed set is closed. Let $K \subseteq Y$ be closed. Then $K_i = f^{-1}(K) \cap C_i = \left(f|_{C_i}\right)^{-1}(K)$ is a closed set in C_i and so is closed in X because C_i is closed. Since $K = K_1 \cup K_2$, K is also closed in X. □

Lemma. Homotopy is an equivalence relation.

Proof. Reflexivity is trivial, because $H(x, t) = f(x)$ is continuous, as $H = f \circ \pi_1$ is the composition of continuous maps. Symmetry holds because if $H(x, t)$ is continuous, $H(x, 1 - t)$ is continuous as the composition of continuous maps. For transitivity, if $f_0 \sim f_1$ via H and $f_1 \sim f_2$ via H', we define

$$H''(x,t) = \begin{cases} H(x, 2t) & t < \frac{1}{2} \\ H'(x, 2t-1) & t \geq \frac{1}{2} \end{cases}$$

and this is continuous by the gluing lemma. □

Note that we sometimes write $f_t(x)$ for a homotopy between f_0 and f_1.

Example. Let $f_1 : X \to \mathbb{R}^n$ be a map. Then $f_0 : X \to \mathbb{R}^n$ defined by $c_{X,0}$ has $f_1 \sim f_0$ via the homotopy $H(x, t) = t f_1(x)$.

Example. Let $f_1 : S^1 \to S^2$ be defined by $f_1(x, y) = (x, y, 0)$: the inclusion map from the circle to the equator in the unit 2-sphere. Let $f_0 : S^1 \to S^2$ be the constant map $f_0(x, y) = (0, 0, 1)$. Then $f_0 \sim f_1$ via the homotopy $f_t(x, y) = (x \sin \frac{\pi t}{2}, y \sin \frac{\pi t}{2}, \cos \frac{\pi t}{2})$.

Lemma. If $f_0, f_1 : X \to Y$ are homotopic via f_t, and $g_0, g_1 : Y \to Z$ are homotopic via g_t, then the map $H : X \times I \to Z$ defined by $H(x, t) = g_t(f_t(x))$, also denoted $g_t \circ f_t$, is a homotopy for $g_0 \circ f_0 \sim g_1 \circ f_1$.

Proof. This is a composition of continuous maps and hence continuous. □

2.2. Contractible spaces

Definition. A space Y is *contractible* if $\mathrm{id}_Y \sim c_{Y,p}$ for some $p \in Y$.

Example. If $Y \subseteq \mathbb{R}^n$ is convex and nonempty, Y is contractible via the homotopy $H(y,t) = (1-t)y + tp$ for some $p \in Y$.

Proposition. Let Y be contractible. Then $f_0 \sim f_1$ for any maps $f_0, f_1 \colon X \to Y$.

Proof. We have $f_0 = \mathrm{id}_Y \circ f_0 \sim c_{Y,p} \circ f_0 = c_{X,p}$, and similarly $f_1 \sim c_{X,p}$. By transitivity, $f_0 \sim f_1$. □

Corollary. Let Y be contractible. Then Y is path-connected.

Proof. If Y is contractible, and $p, q \in Y$, then $c_{\{\cdot\},p} \sim c_{\{\cdot\},q}$ via $H \colon \{\bullet\} \times I \to Y$. Then we can define the path $\gamma(t) = H(\bullet, t)$ from p to q in Y. □

Example. $\mathbb{R} \setminus \{0\}$ is not contractible.

We will later prove that $\mathbb{R}^n \setminus \{0\}$ is not contractible for any $n \geq 1$, but we require some more theory before this can be proven.

Definition. Spaces X, Y are *homotopy equivalent*, denoted $X \sim Y$, if there exist maps $f \colon X \to Y$ and $g \colon Y \to X$ such that $f \circ g \sim \mathrm{id}_Y$ and $g \circ f \sim \mathrm{id}_X$.

Example. If $X \simeq Y$, X and Y are homotopy equivalent. Note that the definition of homotopy equivalence is simply the definition of homeomorphism, except that the requirement that $f \circ g$ and $g \circ f$ be *equal* to the identity is relaxed into simply being *homotopic* to the identity.

Lemma. Homotopy equivalence is an equivalence relation.

Proposition. X is contractible if and only if $X \sim \{\bullet\}$.

Proof. If X is contractible, $\mathrm{id} \sim c_{X,p}$. Let $f \colon X \to \{\bullet\}$ be defined by $f(x) = \bullet$. Let $g \colon \{\bullet\} \to X$ be defined by $g(x) = p$. Then $f \circ g = \mathrm{id}_{\{\cdot\}}$ and $g \circ f = c_{X,p} \sim \mathrm{id}_X$. The converse is similar. □

Example. We have $\mathbb{R}^{n+1} \setminus \{0\} \sim S^n$. Consider $p \colon \mathbb{R}^{n+1} \setminus \{0\} \to S^n$ defined by $p(v) = \frac{v}{\|v\|}$, and $q \colon S^n \to \mathbb{R}^{n+1} \setminus \{0\}$ defined by $q(v) = v$. Then $p \circ q = \mathrm{id}$, and $(q \circ p)(v) = \frac{v}{\|v\|}$. This is homotopic to the identity by

$$H(v,t) = \frac{v}{(1-t) + t\|v\|}$$

This is a special case of a *retract*, a continuous map onto a subspace.

3. Groups from loops

3.1. Homotopy relative to a set

Definition. Let $A \subseteq X$. We say $f_0, f_1 : X \to Y$ are *homotopic relative to* A, written $f_0 \sim f_1 \operatorname{rel} A$, if $f_0 \sim f_1$ via some homotopy $H : X \times I \to Y$ that fixes A, so $H(a,t) = f_0(a) = f_1(a)$ for all $a \in A$.

Lemma. Homotopy relative to A is an equivalence relation.

Lemma. If $f_0, f_1 : X \to Y$ and $f_0 \sim f_1 \operatorname{rel} A$, and $g_0, g_1 : Y \to Z$ and $g_0 \sim g_1 \operatorname{rel} f(A)$, then $g_0 \circ f_0 \sim g_1 \circ f_1 \operatorname{rel} A$.

If $\gamma_0, \gamma_1 : I \to X$ are two homotopic paths relative to their endpoints, so $\gamma_0 \sim \gamma_1 \operatorname{rel} \{0,1\}$, we write $\gamma_0 \sim_e \gamma_1$.

Lemma. Let $f_0, f_1 : I \to I$, where $f_0(0) = f_1(0)$ and $f_0(1) = f_1(1)$. Then $f_0 \sim_e f_1$.

Proof. I is convex, hence $H(x,t) = (1-t)f_0(x) + tf_1(x)$ is a homotopy that preserves endpoints as required. □

Corollary. Suppose $f : I \to I, \gamma : I \to X$. Then if $f(0) = 0$ and $f(1) = 1$, $\gamma \circ f \sim_e \gamma$. Further, if $f(0) = 0$ and $f(1) = 0$, we have $\gamma \circ f \sim_e c_{I,\gamma(0)}$.

Proof. We have $f(0) = \mathrm{id}_I(0)$ and $f(1) = \mathrm{id}_I(1)$. Hence $f \sim_e \mathrm{id}_I$. Therefore, $\gamma \circ f \sim_e \gamma \circ \mathrm{id}_I = \gamma$.

For the second claim, $f(0) = c_{I,0}(0)$ and $f(1) = c_{I,0}(1)$, hence $f \sim_e c_{I,0}$ giving $\gamma \circ f \sim_e \gamma \circ c_{I,0} = c_{I,\gamma(0)}$. □

Definition. Let X be a space, and $p, q \in X$. Let

$$\Omega(X,p,q) = \{\gamma : I \to X \mid \gamma \text{ continuous}, \gamma(0) = p, \gamma(1) = q\}$$

be the set of paths from p to q. Let $\Omega(X,p) = \Omega(X,p,p)$ be the set of loops based at p.

Definition. Let $\gamma \in \Omega(X,p,q), \gamma' \in \Omega(X,q,r)$. Then their composition $\gamma\gamma' \in \Omega(X,p,r)$ is given by

$$(\gamma\gamma')(t) = \begin{cases} \gamma(2t) & t \in \left[0, \frac{1}{2}\right] \\ \gamma'(2t-1) & t \in \left[\frac{1}{2}, 1\right] \end{cases}$$

$\gamma\gamma'$ is continuous by the gluing lemma.

Lemma. Let $\gamma_0, \gamma_1 \in \Omega(X,p,q)$ and $\gamma_0', \gamma_1' \in \Omega(X,q,r)$ such that $\gamma_0 \sim_e \gamma_1$ via $H : I \times I \to X$ and $\gamma_0' \sim_e \gamma_1'$ via $H' : I \times I \to X$. Then $\gamma_0\gamma_0' \sim_e \gamma_1\gamma_1'$.

Proof. The homotopy required is

$$\overline{H}(x,t) = \begin{cases} H(2x,t) & x \in \left[0, \frac{1}{2}\right] \\ H'(2x-1,t) & x \in \left[\frac{1}{2}, 1\right] \end{cases}$$

□

Definition. Let $\gamma \in \Omega(X,p,q)$. Then $\gamma^{-1} \in \Omega(X,q,p)$ is the *reverse* of γ, given by

$$\gamma^{-1}(t) = \gamma(1-t)$$

Proposition. (i) Let $\gamma \in \Omega(X,p,q)$. Then $c_{I,p}\gamma \sim_e \gamma \sim_e \gamma c_{I,q}$.

(ii) $\gamma\gamma^{-1} \sim_e c_{I,p}$ and $\gamma^{-1}\gamma \sim_e c_{I,q}$.

(iii) If $\gamma(1) = \gamma'(0)$ and $\gamma'(1) = \gamma''(0)$, we have

$$\gamma(\gamma'\gamma'') \sim_e (\gamma\gamma')\gamma''$$

Proof. (i) The composition $c_{I,p}\gamma$ has $c_{I,p}\gamma(t) = \gamma(f(t))$ where $f : I \to I$ defined by

$$f(t) = \begin{cases} 0 & t \in \left[0, \frac{1}{2}\right] \\ 2t-1 & t \in \left[\frac{1}{2}, 1\right] \end{cases}$$

Since $f(0) = 0$ and $f(1) = 1$, $\gamma \circ f \sim_e \gamma$. Similarly, $\gamma c_{I,q}(t) = \gamma(g(t))$ where

$$g(t) = \begin{cases} 2t & t \in \left[0, \frac{1}{2}\right] \\ 1 & t \in \left[\frac{1}{2}, 1\right] \end{cases}$$

(ii) $\gamma\gamma^{-1}(t) = \gamma(f(t))$ where

$$f(t) = \begin{cases} 2t & t \in \left[0, \frac{1}{2}\right] \\ 1-2t & t \in \left[\frac{1}{2}, 1\right] \end{cases}$$

Further, $\gamma^{-1}\gamma(t) = \gamma(g(t))$ where

$$g(t) = \begin{cases} 1-2t & t \in \left[0, \frac{1}{2}\right] \\ 2t-1 & t \in \left[\frac{1}{2}, 1\right] \end{cases}$$

(iii) We can write $\gamma(\gamma'\gamma'')(t) = (\gamma\gamma')\gamma(f(t))$ where $f : I \to I$ is the continuous function defined by

$$f(t) = \begin{cases} \frac{t}{2} & t \in \left[0, \frac{1}{2}\right] \\ t - \frac{1}{4} & t \in \left[\frac{1}{2}, \frac{3}{4}\right] \\ 2t-1 & t \in \left[\frac{3}{4}, 1\right] \end{cases}$$

noting that $f(0) = 0$ and $f(1) = 1$. Hence $\gamma(\gamma'\gamma'') \sim_e = (\gamma\gamma')\gamma''$.

□

Definition. Let X be a space and $x_0 \in X$. We define the *fundamental group* or *first homotopy group* of X based at x_0 by

$$\pi_1(X, x_0) = {}^{\Omega(X, x_0)}\big/_{\sim_e}$$

We say x_0 is the *basepoint*. If $\gamma \in \Omega(X, x_0)$, we write $[\gamma]$ for its image in $\pi_1(X, x_0)$, its equivalence class.

Theorem. We define multiplication in π_1 by $[\gamma] * [\gamma'] = [\gamma\gamma']$. The identity is $1 = [c_{I,x_0}]$. The inverse is given by $[\gamma]^{-1} = [\gamma^{-1}]$. These operations form a group.

Proof. Using the above lemma we explicitly check the group axioms. Identity:

$$1[\gamma] = [c_{I,x_0}\gamma] = [\gamma]; \quad [\gamma]1 = [\gamma c_{I,x_0}] = [\gamma]$$

Inverses:

$$[\gamma][\gamma]^{-1} = [\gamma\gamma^{-1}] = [c_{I,x_0}] = 1$$

Associativity:

$$([\gamma][\gamma'])[\gamma''] = [\gamma\gamma'][\gamma''] = [(\gamma\gamma')\gamma''] = [\gamma(\gamma'\gamma'')] = [\gamma][\gamma'\gamma''] = [\gamma]([\gamma'][\gamma''])$$

□

3.2. Induced maps

Definition. Let $f : X \to Y$ be a continuous map, and $f(x_0) = y_0$. Then we have a map $\Omega(X, x_0) \to \Omega(Y, y_0)$ defined by $\gamma \mapsto f \circ \gamma$. Note that if $\gamma_0 \sim_e \gamma_1$, we have $f \circ \gamma_0 \sim_e f \circ \gamma_1$. Thus, this map descends to the *induced homomorphism* $f_\star : \pi_1(X, x_0) \to \pi_1(Y, y_0)$ defined by $[\gamma] \mapsto [f \circ \gamma]$.

Definition. A *pointed space* (X, x_0) is a pair where X is a space and $x_0 \in X$. We write $f : (X, x_0) \to (Y, y_0)$ to denote a map $f : X \to Y$ where $f(x_0) = y_0$. In particular, for $f : (X, x_0) \to (Y, y_0)$ there is an induced map $f_\star : \pi_1(X, x_0) \to \pi_1(Y, y_0)$.

Proposition. Let $f : (X, x_0) \to (Y, y_0)$. Then,

(i) The induced map $f_\star : \pi_1(X, x_0) \to \pi_1(Y, y_0)$ is a group homomorphism.

(ii) $\left(\mathrm{id}_{(X,x_0)}\right)_\star = \mathrm{id}_{\pi_1(X,x_0)}$.

(iii) If $g : (Y, y_0) \to (Z, z_0)$, we have $(g \circ f)_\star = g_\star \circ f_\star$.

(iv) If $f_0, f_1 : (X, x_0) \to (Y, y_0)$ with $f_0 \sim f_1$ rel x_0, then $(f_0)_\star = (f_1)_\star$ (*homotopy invariance*).

Remark. The action of taking the fundamental group of a pointed space thus yields a functor $\pi_1 :$ **Top.** $\to$ **Grp**. The following diagram, representing part (iii) of the proposition above, commutes.

$$\begin{tikzcd}
\pi_1(X, x_0) \arrow[rr, "(g\circ f)_\star"] \arrow[dr, "f_\star"] & & \pi_1(Z, z_0) \\
& \pi_1(Y, y_0) \arrow[ur, "g_\star"] & \\
& (Y, y_0) \arrow[u, dashed] \arrow[dr, "g"] & \\
(X, x_0) \arrow[uuu, dashed] \arrow[ur, "f"] \arrow[rr, "g\circ f"] & & (Z, z_0) \arrow[uuu, dashed]
\end{tikzcd}$$

Proof. (i) This follows from the fact that

$$f \circ (\gamma\gamma')(t) = \begin{cases} f \circ \gamma(2t) & t \in \left[0, \frac{1}{2}\right] \\ f \circ \gamma'(2t-1) & t \in \left[\frac{1}{2}, 1\right] \end{cases} = (f \circ \gamma)(f \circ \gamma')(t)$$

Hence,

$$f_\star([\gamma][\gamma']) = [f \circ (\gamma\gamma')] = [(f \circ \gamma)(f \circ \gamma')] = [f \circ \gamma][f \circ \gamma'] = f_\star([\gamma])f_\star([\gamma'])$$

(ii) $\mathrm{id}_\star([\gamma]) = [\mathrm{id}_X \circ \gamma] = [\gamma]$.

(iii) $(f \circ g)_\star([\gamma]) = [f \circ g \circ \gamma] = f_\star([g \circ \gamma]) = f_\star(g_\star([\gamma]))$.

(iv) $f_0 \sim f_1 \operatorname{rel} x_0$ and $\gamma(0) = \gamma(1) = x_0$ implies $f_0 \circ \gamma \sim_e f_1 \circ \gamma$, so $(f_0)_\star([\gamma]) = (f_1)_\star([\gamma])$.

□

Example. Let $f : X \to Y$ be a homeomorphism, and let $y_0 = f(x_0)$. Then $f : (X, x_0) \to (Y, y_0)$ and $f^{-1} : (Y, y_0) \to (X, x_0)$ are inverses. Thus, $f_\star : \pi_1(X, x_0) \to \pi_1(Y, y_0)$ and $f_\star^{-1} : \pi_1(Y, y_0) \to \pi_1(X, x_0)$ are inverses. Since $f_\star \circ f_\star^{-1} = (f \circ f^{-1})_\star = \mathrm{id}_{\pi_1(Y, y_0)}$ and $f_\star^{-1} \circ f_\star = \mathrm{id}_{\pi_1(X, x_0)}$, we have that $f_\star$ is a group isomorphism, and π_1 is a topological invariant.

3.3. Retractions

Definition. Let $A \subset X$, where $\iota \colon A \to X$ is the inclusion map. Then $p \colon X \to A$ is a *retraction* if $p \circ \iota = \mathrm{id}_A$. $p \colon X \to A$ is a *strong deformation retraction*, or *s.d.r.*, if $p \circ \iota = \mathrm{id}_A$ and $\iota \circ p \sim \mathrm{id}_X \operatorname{rel} A$.

Remark. In either case, if $a_0 \in A$, $\iota \colon (A, a_0) \to (X, a_0)$ and $p \colon (X, a_0) \to (A, a_0)$. If p is a retraction, $p_\star \circ \iota_\star = (p \circ \iota)_\star = (\mathrm{id}_A)_\star = \mathrm{id}_{\pi_1(A, a_0)}$, so $\iota_\star : \pi_1(A, a_0) \to \pi_1(X, a_0)$ is injective, and $p_\star : \pi_1(X, a_0) \to \pi_1(A, a_0)$ is surjective. If p is a strong deformation retraction, $\iota_\star \circ p_\star = (\iota \circ p)_\star = (\mathrm{id}_X)_\star = \mathrm{id}_{\pi_1(X, a_0)}$, so $p_\star$ and $\iota_\star$ are isomorphisms.

Remark. If $p \colon X \to A$ is a strong deformation retraction, then $A \sim X$.

Example. $p : \mathbb{R}^{n+1} \setminus \{0\} \to S^n$ given by $v \mapsto \frac{v}{\|v\|}$ is a strong deformation retraction.

Example. $\mathbb{R}^2 \setminus \{0, 1\}$ has A, B as strong deformation retractions, where A is a figure-eight with one loop surrounding each hole, and B is a rectangle surrounding each hole with a vertical line connecting the top and bottom edges through $\left(\frac{1}{2}, 0\right)$. This can be a useful trick to show $A \sim B$.

3.4. Null-homotopy and extensions

Definition. We say $f : X \to Y$ is *null-homotopic* if $f \sim c_{X,p}$ for $p \in Y$.

Example. If X is contractible, then $\mathrm{id}_X \sim c_{X,q}$, so $f = f \circ \mathrm{id}_X \sim f \circ c_{X,q} = f(q)$. So any $f : X \to Y$ is null-homotopic. If $f_0 \sim f_1$, then f_0 is null-homotopic if and only if f_1 is null-homotopic.

Definition. Let $A \subset X$ and $f : A \to Y$. We say a continuous map $F : X \to Y$ is an *extension* of f if $F|_A = f$. If such a map exists, we say f *extends* to X.

$$\begin{tikzcd} & X \arrow[d, dashed, "F"] \\ A \arrow[ur, "\iota"] \arrow[r, "f"'] & Y \end{tikzcd}$$

Lemma. $f : S^1 \to Y$ extends to D^2 if and only if f is null-homotopic.

Proof. If F is an extension of f to D^2, we define $H(v, t) = F(tv)$. Then H is a homotopy from f to $c_{S^1,F(0)}$. So f is null-homotopic.

Conversely, if f is null-homotopic, let $H : S^1 \times I \to Y$ be a homotopy for $c_{S^1,p} \sim f$. Then we define

$$F(v) = \begin{cases} H\left(\frac{v}{\|v\|}, \|v\|\right) & v \neq 0 \\ p & v = 0 \end{cases}$$

One can check that this is indeed a continuous extension. □

Definition. Let $\gamma \in \Omega(X, x_0)$. We define $\overline{\gamma} : S^1 \to X$ by $\overline{\gamma}(e^{2\pi i t}) = \gamma(t)$. This is well-defined since $\gamma(0) = \gamma(1)$, and it is continuous because $I/\{0,1\} \simeq S^1$.

Lemma. (i) If $\gamma_0 \sim_e \gamma_1$ via $H(x, t)$, we have $\overline{\gamma}_0 \sim \overline{\gamma}_1$ via $\overline{H} : S^1 \times I \to Y$ given by $\overline{H}(e^{2\pi i x}, t) = H(x, t)$.

(ii) $\overline{\gamma\gamma'} \sim \overline{\gamma'\gamma}$.

Proof. (i) Note that $\overline{H}$ is well-defined since $H(0, t) = H(1, t) = x_0$.

(ii) We have $\overline{\gamma\gamma'}(v) = \overline{\gamma'\gamma}(-v)$, hence $\overline{\gamma\gamma'} = \overline{\gamma'\gamma} \circ a$ where $a : S^1 \to S^1$ is the antipodal map. Since $a \sim \mathrm{id}_{S^1}$, we have $\overline{\gamma\gamma'} \sim \overline{\gamma'\gamma}$.

□

Consider the radial projection homeomorphism $\Phi : D^2 \to I \times I$. Note that $\Phi(S^1) = \partial(I \times I) = I \times \{0,1\} \cup \{0,1\} \times I$. Since Φ is a homeomorphism, $h : \partial(I \times I) \to X$ extends to $I \times I$ if and only if $h \circ \Phi$ extends to D^2, which is true if and only if $h \circ \Phi$ is null-homotopic. Define $\alpha_i(t) = h(t,i)$ and $\beta_i(t) = h(i,t)$ for $i = 0,1$. Then, $h \circ \Phi \sim \overline{\alpha_0\beta_1\alpha_1^{-1}\beta_0^{-1}}$.

Proposition. Let $\gamma_0, \gamma_1 \in \Omega(X, p, q)$. Then the following are equivalent.

(i) $\gamma_0 \sim_e \gamma_1$;

(ii) $\overline{\gamma_0\gamma_1^{-1}}$ is null-homotopic;

(iii) $[\gamma_0\gamma_1^{-1}] = 1$ in $\pi_1(X, p)$.

Proof. Consider $h : \partial(I \times I) \to X$ given by $\gamma_0 c_{I,q} \gamma_1^{-1} c_{I,p}$. Note that h is continuous by the gluing lemma. $\gamma_0 \sim_e \gamma_1$ if and only if h extends to $I \times I$, which is true if and only if $h \circ \Phi$ extends to D^2, if and only if $\overline{\gamma_0 c_{I,q} \gamma_1^{-1} c_{I,p}}$ is null-homotopic. But this is homotopic to $\overline{\gamma_0\gamma^{-1}}$, so this proves that (i) and (ii) are equivalent.

Now, consider $h' : \partial(I \times I) \to X$ given by $\gamma_0\gamma_1^{-1}$ on one side, and on all other sides, $c_{I,p}$. Then $[\gamma_0\gamma_1^{-1}] = 1$ if and only if $\gamma_0\gamma_1^{-1} \sim_e c_{I,p}$, if and only if h' extends to $I \times I$, if and only if $h \circ \Phi$ extends to D^2, if and only if $\overline{\gamma_0\gamma_1^{-1} c_{I,p} c_{I,p}^{-1} c_{I,p}^{-1}} \sim \overline{\gamma_0\gamma_1^{-1}}$ is null-homotopic. □

Corollary. The following are equivalent.

(i) $\gamma_0 \sim_e \gamma_1$ for all $\gamma_0, \gamma_1 \in \Omega(X, p, q)$ and all $p, q \in X$.

(ii) any $f : S^1 \to X$ is null-homotopic;

(iii) $\pi_1(X, x_0)$ is the trivial group for all $x_0 \in X$.

Definition. X is *simply connected* if X is path-connected and $\pi_1(X, x_0) = 1$ for all $x_0 \in X$.

3.5. Change of basepoint

Lemma. Let X_0 be the path-connected component of X containing a point $x_0 \in X$. If Z is path-connected, $f : Z \to X$ is continuous, and $x_0 \in \operatorname{Im} f$, we have $\operatorname{Im} f \subseteq X_0$.

Proof. Suppose $f(z_0) = x_0$. Given $z \in Z$, choose $\gamma \in \Omega(Z, z_0, z)$ by path-connectedness. Then $f \circ \gamma \in \Omega(X, x_0, f(z))$, so $f(Z) \subseteq X_0$. □

Let $\iota : (X_0, x_0) \to (X, x_0)$ be the inclusion map. Then if $f : (Z, z_0) \to (X, x_0)$ and Z is path-connected, f factors through ι as $f = \iota \circ \hat{f}$ where $\hat{f} : (Z, z_0) \to (X_0, x_0)$.

Lemma. The map $\iota_\star : \pi_1(X_0, x_0) \to \pi_1(X, x_0)$ is an isomorphism.

Proof. Let $[\gamma] \in \pi_1(X, x_0)$, so $\gamma : (I, 0) \to (X, x_0)$ giving $\gamma = \iota \circ \hat{\gamma}$ where $\hat{\gamma} \in \Omega(X_0, x_0)$; $[\gamma] = \iota_\star([\hat{\gamma}])$, so $\iota_\star$ is surjective. Now suppose $\gamma_0 = \iota \circ \hat{\gamma}_0, \gamma_1 = \iota \circ \hat{\gamma}_1$. If $\iota_\star([\hat{\gamma}_0]) = \iota_\star([\hat{\gamma}_1])$, so $\gamma_0 \sim_e \gamma_1$ via $H : I \times I \to X$, we have $H(0, 0) = x_0$, so $H = \iota \circ \hat{H}$ since $I \times I$ is path-connected. Then we can check $\hat{H}$ is a homotopy for $\hat{\gamma}_0 \sim_e \hat{\gamma}_1$. Hence $[\hat{\gamma}_0] = [\hat{\gamma}_1]$, so $\iota_\star$ is injective. □

Let $u \in \Omega(X, x_0, x_1)$. Then we can define $u_\sharp : \Omega(X, x_0) \to \Omega(X, x_1)$ by $\gamma \mapsto u^{-1}\gamma u$. Hence if $\gamma_0 \sim_e \gamma_1$, we have $u^{-1}\gamma_0 u \sim_e u^{-1}\gamma_1 u$, so $u_\sharp$ descends to a map $u_\sharp : \pi_1(X, x_0) \to \pi_1(X, x_1)$ defined by $[\gamma] \mapsto [u^{-1}\gamma u]$.

Proposition. $u_\sharp$ is a group isomorphism with inverse $(u^{-1})_\sharp$.

$$\begin{tikzcd}
\pi_1(X, x_0) \arrow[r, shift left, "u_\sharp"] & \pi_1(X, x_1) \arrow[l, shift left, "(u^{-1})_\sharp"]
\end{tikzcd}$$

Proof. First, it is a homomorphism.

$$\begin{aligned}
u_\sharp([\gamma][\gamma']) &= [u^{-1}\gamma\gamma' u] = [u^{-1}\gamma c_{I,x_0}\gamma' u] \\
&= [u^{-1}\gamma u u^{-1}\gamma' u] = [u^{-1}\gamma u][u^{-1}\gamma' u] = u_\sharp([\gamma])u_\sharp([\gamma'])
\end{aligned}$$

Consider the function $u_\sharp^{-1}$. We have

$$u_\sharp^{-1}(u_\sharp([\gamma])) = [uu^{-1}\gamma uu^{-1}] = [c_{I,x_0}\gamma c_{I,x_0}] = [\gamma]$$

and

$$u_\sharp(u_\sharp^{-1}([\gamma])) = [u^{-1}u\gamma u^{-1}u] = [c_{I,x_1}\gamma c_{I,x_1}] = [\gamma]$$

So $u_\sharp, u_\sharp^{-1}$ are inverses, and therefore isomorphisms. □

Corollary. A space X is simply connected if it is path-connected and $\pi_1(X, x_0) = 1$ for any $x_0 \in X$, since then it follows that $\pi_1(X, x) = 1$ for all $x \in X$.

Theorem. Let $x_0 \in X$, and $f_0, f_1 : X \to Y$ such that $f_0 \sim f_1$ by $H : X \times I \to Y$. Let $u(t) = H(x_0, t)$ and $y_0 = f_0(x_0), y_1 = f_1(x_0)$. Then $u \in \Omega(Y, y_0, y_1)$. We have $f_i : (X, x_0) \to (Y, y_i)$ which induce $f_{i\star} : \pi_1(X, x_0) \to \pi_1(Y, y_i)$. Then $f_{1\star} = u_\sharp \circ f_{0\star}$.

$$\begin{tikzcd}
 & (Y, y_0) \\
(X, x_0) \arrow[ur, "f_0"] \arrow[dr, "f_1"'] & \\
 & (Y, y_1)
\end{tikzcd}
\qquad
\begin{tikzcd}
 & \pi_1(Y, y_0) \arrow[dd, "u_\sharp"] \\
\pi_1(X, x_0) \arrow[ur, "f_{0\star}"] \arrow[dr, "f_{1\star}"] & \\
 & \pi_1(Y, y_1)
\end{tikzcd}$$

Proof. We must show that $f_{1\star}([\gamma]) = u_\sharp(f_{0\star}([\gamma]))$. Let $\gamma_i = f_i \circ \gamma$. We therefore need to show $\gamma_1 \sim_e u^{-1}\gamma_0 u$ for all $\gamma \in \Omega(X, x_0)$. Suppose we can show that $H : \partial(I \times I) \to Y$ given by $\gamma_0, u, \gamma_1^{-1}, u^{-1}$ on each side of the square extends to $I \times I$. Equivalently, $\overline{\gamma_0 u \gamma_1^{-1} u^{-1}} = u^{-1}\gamma_0 u \gamma_1^{-1}$ is null-homotopic. This is equivalent to the statement $u^{-1}\gamma_0 u \sim_e \gamma_1$. We know h extends to $\hat{H} : I \times I \to Y$, because $\hat{H}(x,t) = H(\gamma(x), t)$. □

Corollary. Let $X \sim Y$ via $f : X \to Y$ and $g : Y \to X$, so $f \circ g \sim \mathrm{id}_Y$ and $g \circ f \sim \mathrm{id}_X$. Let $x_0 \in X$ and $f(x_0) = y_0$. Let $g(y_0) = x_1$ and $f(x_1) = y_1$. Then we have induced maps $f_\star^{(0)} : \pi_1(X, x_0) \to \pi_1(Y, y_0)$, $g_\star : \pi_1(Y, y_0) \to \pi_1(X, x_1)$, $f_\star^{(1)} : \pi_1(X, x_1) \to \pi_1(Y, y_1)$. Then $g_\star$ is an isomorphism.

$$\begin{tikzcd}
(X,x_0) \arrow[dr, "f"] \arrow[dd, dashed, "g\circ f\sim \mathrm{id}_X"'] & \\
& (Y,y_0) \arrow[dl, "g"] \arrow[dd, dashed, "f\circ g\sim \mathrm{id}_Y"] \\
(X,x_1) \arrow[dr, "f"'] & \\
& (Y,y_1)
\end{tikzcd}
\qquad
\begin{tikzcd}
\pi_1(X,x_0) \arrow[dr, "f_\star^{(0)}"] \arrow[dd, "u_\sharp"'] & \\
& \pi_1(Y,y_0) \arrow[dl, "g_\star"] \arrow[dd, "u_\sharp"] \\
\pi_1(X,x_1) \arrow[dr, "f_\star^{(1)}"] & \\
& \pi_1(Y,y_1)
\end{tikzcd}$$

The left-hand commutative diagram, in the category of pointed topological spaces, commutes up to homotopy. The right-hand induced diagram commutes.

Proof. We have $\mathrm{id}_X \sim g \circ f$ via $H : X \times I \to X$. Then $g_\star \circ f_\star^{(0)} = (g \circ f)_\star = u_\sharp \circ (\mathrm{id}_X)_\star$ where $u(t) = H(x_0, t)$ is a path from x_0 to x_1. Since $u_\sharp$ is an isomorphism, $g_\star$ is surjective. Similarly, $f^{(1)} \circ g_\star = (f \circ g)_\star$ is an isomorphism, so $g_\star$ is injective. □

Corollary. Let X be contractible. Then $\pi_1(X, x_0) = 1$ is the trivial group.

Proof. The space $\Omega(\{\bullet\}, \bullet)$ has one element, so $\pi_1(\{\bullet\}, \bullet) = 1$. Since $X \sim \{\bullet\}$, the result follows. □

4. Covering spaces

4.1. Definitions

Definition. Let $p : \hat{X} \to X$ be a continuous function. We say $U \subset X$ is *evenly covered* by p if $p^{-1}(U) \simeq \coprod_{\alpha \in A} U_\alpha$ and $p|_{U_\alpha} : U_\alpha \to U$ is a homeomorphism for all α.

The topology on the coproduct $\coprod_{\alpha \in A} U_\alpha$ is such that V is open if and only if each projection $V \cap U_\alpha$ is open. The topology on $p^{-1}(U)$ is the subspace topology. In particular, the inclusions $\iota_\alpha : U_\alpha \to \coprod_{\alpha \in A} U_\alpha \to \hat{X}$ are continuous, as is the composition $\iota_\alpha\big(p|_{U_\alpha}\big)^{-1} : U \to X$ since $p|_{U_\alpha}$ is a homeomorphism.

Definition. $p : \hat{X} \to X$ is a *covering map* if every $x \in X$ has an open neighbourhood U_x which is evenly covered by p. If so, we say $\hat{X}$ is a *covering space* of X.

Example. If A is a space with the discrete topology, then $p : A \times X \to X$ is a covering map, because $p^{-1}(X) = \coprod_{\alpha \in A} \{\alpha\} \times X$.

Example. $p : \mathbb{R} \to S^1$ given by $p(t) = e^{2\pi i t}$ is a covering map. Indeed, if $V \subseteq \mathbb{R}$ is an open interval of at most unit length, let $U = p(V)$ and then $p^{-1}(U) = \coprod_{n \in \mathbb{Z}} V_n$ for $V_n = \{n + v \mid v \in V\}$.

Example. Consider $p_n : S^1 \to S^1$ defined by $z \mapsto z^n$. If $V \subseteq S^1$ is an open interval of length $< \frac{2\pi}{n}$, let $U = p_n(V)$. Then $p_n^{-1}(V) = \coprod_{i \in \mathbb{Z}/n\mathbb{Z}} \omega^i V$ for $\omega = e^{\frac{2\pi i}{n}}$. Hence U is evenly covered.

Definition. We define the n-dimensional real projective space as $\mathbb{RP}^n = S^n/\sim$ where $\sim$ is the equivalence relation generated by $x \sim -x$ for all $x \in S^n$.

Example. The quotient map $p : S^n \to \mathbb{RP}^n$ is a covering map. Indeed, for $x \in S^n$, let V_x be the open hemisphere centred at x. Then letting $U_x = p(V_x)$, we have $p^{-1}(U(x)) = U_x \amalg -U_x$, giving that U_x is evenly covered.

4.2. Lifting paths and homotopies

Definition. Let $p : \hat{X} \to X$ be a covering map, and $f : Z \to X$ be continuous. A continuous function $\hat{f} : Z \to \hat{X}$ is a *lift* if $p \circ \hat{f} = f$. Hence, the following commutative diagram holds.

$$\begin{tikzcd}
 & \hat{X} \arrow[d, "p"] \\
Z \arrow[ur, dashed, "\hat{f}"] \arrow[r, "f"'] & X
\end{tikzcd}$$

Theorem (Path lifting). Let $p : (\hat{X}, \hat{x}_0) \to (X, x_0)$ be a covering map, and $\gamma : [a, b] \to X$ be a path. Let $\gamma(a) = x_0$ and $p(\hat{x}_0) = \hat{x}_0$. Then there exists a unique lift $\hat{\gamma} : [a, b] \to \hat{X}$ with $\hat{\gamma}(a) = \hat{x}_0$.

The proof will be given after some lemmas. We say $f : Z \to X$ has the *(unique) lifting property at $z \in Z$* if for any $\hat{x} \in \hat{X}$ such that $p(\hat{x}) = f(z)$, there exists a (unique) lift $\hat{f} : Z \to \hat{X}$ such that $\hat{f}(z) = \hat{x}$.

Lemma (Lebesgue covering lemma)**.** Let X be a compact metric space, and $\{U_\alpha \mid \alpha \in A\}$ is an open cover of X. Then there exists $\delta > 0$ such that for every $x \in X$, the open ball $B_\delta(x)$ is contained in U_α for some $\alpha \in A$.

Proof. We have an open cover $\{U_\alpha \mid \alpha \in A\}$ of X, so given $x \in X$, we can find $\alpha_x \in A$ such that $x \in U_{\alpha_x}$ and U_{α_x} is open. Hence there exists $\delta_x > 0$ such that $B_{2\delta_x}(x) \subset U_{\alpha_x}$. Then $\left\{B_{\delta_x}(x) \mid x \in X\right\}$ is an open cover of X. By compactness there is a finite subcover $\left\{B_{\delta_{x_i}}(x_i) \mid i \in \{1, \dots, k\}\right\}$. Let $\delta = \min_{i \in \{1,\dots,k\}} \delta_{x_i} > 0$. Then for $y \in X$, we have $y \in B_{\delta_{x_i}}(x_i)$ for some i, and $B_\delta(y) \subset B_{\delta_{x_i}+\delta}(x_i) \subset B_{2\delta_{x_i}}(x_i) \subset U_{\alpha_x}$. □

Lemma. Let $p : (\hat{X}, \hat{x}_0) \to (X, x_0)$ be a covering map, and $\gamma : [a, b] \to X$ be a path such that $\gamma(a) = x_0$. Let $\operatorname{Im}\gamma \subset U$ where $U \subset X$ is evenly covered. Then γ has the unique lifting property.

Note that this is simply the above path lifting theorem with an additional hypothesis.

Proof. Since U is evenly covered, $p^{-1}(U) = \coprod_{\alpha \in A} U_\alpha$, and $p|_{U_\alpha} : U_\alpha \to U$ is a homeomorphism onto its image. So $\hat{x}_0 \in U_{\alpha_0}$ for some $\alpha_0 \in A$. Then the map $(p_\alpha)^{-1} = \iota_\alpha \circ \left(p|_{U_\alpha}\right)^{-1} : U \to \hat{X}$ is continuous. Then $\left(p|_{U_{\alpha_0}}\right)^{-1}(x_0) = \hat{x}_0$, so $\hat{\gamma} = \left(p_{\alpha_0}\right)^{-1} \circ \gamma$ is a lift of γ with $\hat{\gamma}(a) = \hat{x}_0$.

Now we will prove uniqueness of the lift. Observe that $p^{-1}(U) = U_{\alpha_0} \amalg \coprod_{\alpha \neq \alpha_0} U_\alpha$ disconnects $p^{-1}(U)$. Note that $[a, b]$ is connected. We have that if $\hat{\gamma} : [a, b] \to \hat{X}$ with $\hat{\gamma}(a) = \hat{x}_0$ and $p \circ \hat{\gamma} = \gamma$, then $\operatorname{Im}\hat{\gamma} \subset p^{-1}(U)$ implies $\operatorname{Im}\hat{\gamma} \subset U_{\alpha_0}$. But $p|_{U_{\alpha_0}}$ is a homeomorphism, so we must have $\hat{\gamma} = \left(p_{\alpha_0}\right)^{-1} \circ \gamma$. □

Lemma. Let $\gamma : [a, b] \to X$ and $a' \in [a, b]$. If $\gamma|_{[a,a']}$ has the unique lifting property at a and $\gamma|_{[a',b]}$ has the unique lifting property at a', then γ has the unique lifting property at a.

Proof. If $p(\hat{x}) = \gamma(a)$, since $\gamma|_{[a,a']}$ has the unique lifting property at a, there exists a unique lift $\hat{\gamma}_1 : [a, a'] \to \hat{X}$ such that $\hat{\gamma}_1(a) = \hat{x}$. Then $\gamma|_{[a',b]}$ has the unique lifting property at a', so there exists a unique lift $\hat{\gamma}_2 : [a', b] \to \hat{X}$ with $\hat{\gamma}_2(a') = \hat{\gamma}_1(a')$. Then the composition $\hat{\gamma} = \hat{\gamma}_1\hat{\gamma}_2$ is a lift of γ, with $\hat{\gamma}(a) = \hat{x}$.

For uniqueness, suppose $\hat{\gamma}$ is a lift of γ with $\hat{\gamma}(a) = \hat{x}$. Then $\hat{\gamma}|_{[a,a']}$ is a lift of $\gamma|_{[a,a']}$, so by the unique lifting property, $\hat{\gamma}|_{[a,a']}$ is uniquely determined such that $\hat{\gamma}(a) = \hat{x}$. Then by the unique lifting property again, $\hat{\gamma}|_{[a',b]}$ is also uniquely determined such that $\hat{\gamma}|_{[a',b]}(a') = \hat{\gamma}|_{[a,a']}(a')$. □

We can now prove the path lifting theorem: any $\gamma : I \to X$ has the unique lifting property.

Proof. Let $p : \hat{X} \to X$ be a covering map. Hence, for all $x \in X$, there exists an open neighbourhood U_x which is evenly covered. $\{U_x \mid x \in X\}$ is therefore an open cover of X, and so $\{\gamma^{-1}(U_x) \mid x \in X\}$ is an open cover of I. Since I is compact, by the Lebesgue covering lemma, there exists $\delta > 0$ such that for all t, $B_\delta(t) \subseteq \gamma^{-1}(U_{x(t)})$ for some $x(t)$. In other words, $\gamma(B_\delta(t)) \subseteq U_{x(t)}$.

Let $n \in \mathbb{N}$ such that $\frac{1}{n} < \delta$, and $a_i = \frac{i}{n} \in I$. Then $[a_i, a_{i+1}] \subset B_\delta(a_i)$ for all i. Hence $\gamma[a_i, a_{i+1}] \subseteq U_{x(a_i)}$. Then $[a_i, a_{i+1}]$ is connected, hence $\gamma[a_i, a_{i+1}]$ is connected. Since $U_{x(a_i)}$ is evenly covered, $\gamma|_{[a_i,a_{i+1}]}$ has the unique lifting property. Then by induction on i, we can see that $\gamma|_{[0,a_i]}$ has the unique lifting property, and hence so does γ in its entirety. □

Theorem (Homotopy lifting). Let $p : (\hat{X}, \hat{x}_0) \to (X, x_0)$ be a covering map, and $H : I \times I \to X$ be a homotopy. Then H has the lifting property at $(0, 0)$.

It also has the unique lifting property, but this will be more easily proven later.

Proof. I is compact and connected, so by Tychonoff's theorem, $I \times I$ is compact and connected. Suppose $\{U_x \mid x \in X\}$ is an open cover of X consisting of evenly covered neighbourhoods of points as before. Then, since $I \times I$ is compact, by the Lebesgue covering lemma there exists $\delta > 0$ such that for all $v \in I \times I$, $B_\delta(v) \subseteq H^{-1}(U_{x(v)})$. In particular, $H(B_\delta(v)) \subseteq U_{x(v)}$.

Let $n \in \mathbb{N}$ such that $\frac{\sqrt{2}}{n} < \delta$, dividing $I \times I$ into squares of size $\frac{1}{n}$, ordered from left-to-right and then bottom-to-top. Label each square with an index $i \in \{1, \dots, n^2\}$. Let each square A_i have lower left-hand corner v_i, for $i \in \{1, \dots, n^2\}$. Note that $H(A_i) \subseteq H(B_\delta(v_i)) \subseteq U_{x(v_i)} = U_i$ is evenly covered.

Let $B_k = \bigcup_{i=1}^k A_i$. Then $A_i \simeq I \times I$ is connected, so $H|_{A_i}$ has the lifting property at v_i.

We show by induction that $H|_{B_k}$ has the lifting property at $(0, 0)$. For $k = 1$, $B_1 = A_1$ and $(0, 0) = v_1$, so the result follows.

For other k, suppose that $H|_{B_k}$ has the lifting property at $(0, 0)$, so $\hat{H}_k : B_k \to \hat{X}$ with $\hat{H}_k(0, 0) = \hat{x}$. Then $H|_{A_{k+1}}$ has the lifting property at v_i, so choose a lift $\hat{h}_k : A_{k+1} \to \hat{X}$ such that $\hat{h}_k(v_{k+1}) = \hat{H}_k(v_{k+1})$. Note that $p(\hat{H}_k(v_{k+1})) = H(v_{k+1})$, so this exists by the lifting property. Observe that $A_{k+1} \cap B_k = I_k \cup I'_k$ is the union of (at most) two intervals with intersection at their endpoints, so is homeomorphic to I. Hence by uniqueness of path lifting, $\hat{H}_k|_{I_k} = \hat{h}_k|_{I_k}$ since both are lifts of $H|_{I_k}$ with $v_{k+1} \mapsto \hat{H}_k(v_{k+1})$. Similarly, $\hat{H}_k|_{I'_k} = \hat{h}_k|_{I'_k}$. In other words, $\hat{H}_k|_{A_{k+1} \cap B_k} = \hat{h}_k|_{A_{k+1} \cap B_k}$. By the gluing lemma, we can construct the well-defined and continuous map $\hat{H}_{k+1} : B_{k+1} \to X$ given by $\hat{H}_k$ and $\hat{h}_k$ on their domains. Then $\hat{H}_{k+1}$ is a lift of $H|_{B_{k+1}}$. □

Proposition. Let $p\colon (\hat{X}, \hat{x}_0) \to (X, x_0)$ be a covering map. Let $\gamma_0, \gamma_1 \in \Omega(X, x_0, x_1)$, and $\gamma_0 \sim_e \gamma_1$. Let $\hat{\gamma}_i$ be the lift of γ_i to $\hat{X}$ with $\hat{\gamma}_i(0) = \hat{x}_0$, which exists by the path lifting property. Then $\hat{\gamma}_0 \sim_e \hat{\gamma}_1$.

Proof. Let $H\colon I \times I \to X$ be a homotopy between γ_0 and γ_1. By the homotopy lifting property, there exists a lifted homotopy $\hat{H}\colon I \times I \to \hat{X}$ such that $\hat{H}(0,0) = \hat{x}_0$. Let $\alpha_i(t) = \hat{H}(t,i)$ for $i = 0,1$, and $\beta_i(t) = \hat{H}(i,t)$ for $i = 0,1$. Applying the uniqueness of path lifting to the α_i and the β_i,

(i) α_0 is a lift of γ_0 with $\alpha_0(0) = \hat{x}_0$, so $\alpha_0 = \hat{\gamma}_0$;

(ii) β_0 is a lift of c_{I,x_0} with $\beta_0(0) = \hat{x}_0$, so $\beta_0 = \hat{c}_{I,x_0} = c_{I,\hat{x}_0}$ by uniqueness, and in particular, $\alpha_1(0) = \beta_0(1) = \hat{x}_0$;

(iii) α_1 is a lift of γ_1 with $\alpha_1(0) = \hat{x}_0$, so $\alpha_1 = \hat{\gamma}_1$;

(iv) let $\hat{x}_1 = \hat{\gamma}_0(1)$, and then β_1 is a lift of c_{I,x_1}, so $\beta_1(0) = \hat{x}_1$, so $\beta_1 = c_{I,\hat{x}_1}$.

Hence $\hat{\gamma}_0 \sim_e \hat{\gamma}_1$ via $\hat{H}$. □

Corollary. Let $p\colon (\hat{X}, \hat{x}_0) \to (X, x_0)$ be a covering map. Let $\gamma_0, \gamma_1 \in \Omega(X, x_0, x_1)$, and $\gamma_0 \sim_e \gamma_1$. Then $\hat{\gamma}_0(1) = \hat{\gamma}_1(1)$.

4.3. Simply connected lifting

Let $p\colon (\hat{X}, \hat{x}_0) \to (X, x_0)$ be a covering map. If $\gamma\colon I \to X$ has $\gamma(0) = x_0$, let $\hat{\gamma}\colon I \to \hat{X}$ be its unique lift such that $\hat{\gamma}(0) = \hat{x}_0$.

Consider $\widehat{\gamma\gamma'} = \hat{\gamma}\tilde{\gamma}'$, where $\tilde{\gamma}'$ is a lift of γ' such that $\tilde{\gamma}'(0) = \hat{\gamma}(1)$. Note that we needed to change the start point of $\tilde{\gamma}'$ in the covering space.

Definition. A space X is *locally path-connected* if for every open set $U \subseteq X$ and $x \in U$, there exists an open $V \subseteq U$ with $x \in V$ and V path-connected.

Example. Consider

$$X = \left\{(x,0) \in \mathbb{R}^2\right\} \cup \left\{\left(\frac{1}{n}, y\right) \in \mathbb{R}^2, n \in \mathbb{Z}\right\} \cup \left\{(0,y) \in \mathbb{R}^2\right\}$$

Then, an open set containing a point $(0,y)$ but not $(0,0)$ admits no smaller path-connected open neighbourhood.

Proposition (simply connected lifting property)**.** Let Z be a simply connected (and hence path-connected) space that is also locally path-connected. If $f\colon (Z, z_0) \to (X, x_0)$, then f has a unique lift $\hat{f}\colon (Z, z_0) \to (\hat{X}, \hat{x}_0)$.

Remark. This proposition then implies the path lifting and homotopy lifting properties.

Proof. Suppose $\hat{f} : (Z, z_0) \to (\hat{X}, \hat{x}_0)$ is a lift of f. Given $z \in Z$, consider a path $\gamma \in \Omega(Z, z_0, z)$, which exists since Z is path-connected. Then $\hat{f} \circ \gamma$ is a lift of $f \circ \gamma$, since $p(\hat{f} \circ \gamma) = (p \circ \hat{f}) \circ \gamma = f \circ \gamma$. Then, $(\hat{f} \circ \gamma)(0) = \hat{f}(z_0) = \hat{x}_0$, so $\hat{f} \circ \gamma = \widehat{f \circ \gamma}$ is the unique lift of $f \circ \gamma$ given by the unique path lifting property. Then $\hat{f}(z) = \hat{f}(\gamma(1)) = (\hat{f} \circ \gamma)(1) = \widehat{f \circ \gamma}(1)$ is uniquely determined by the unique path lifting property. So any such lift is unique.

If $\gamma_0, \gamma_1 \in \Omega(Z, z_0, z)$, $\gamma_0 \sim_e \gamma_1$ by simply-connectedness. In particular, $f \circ \gamma_0 \sim_e f \circ \gamma_1$, and by the homotopy lifting property, $\widehat{f \circ \gamma_0}(1) = \widehat{f \circ \gamma_1}(1)$. So the choice of path γ used above is not relevant. Now, let us define $\hat{f} : (Z, z_0) \to (\hat{X}, \hat{x}_0)$ by $\hat{f}(z) = \widehat{f \circ \gamma}(1)$ where $\gamma \in \Omega(Z, z_0, z)$ is any path from z_0 to z. Then $p(\hat{f}(z)) = p \circ \widehat{f \circ \gamma}(1) = f \circ \gamma(1) = f(z)$ since $\widehat{f \circ \gamma}$ is a lift of $f \circ \gamma$. Hence $\hat{f}$ as defined is a lift. If $z = z_0$, we can take $\gamma = c_{I, z_0}$, so $f \circ \gamma = c_{I, x_0}$. In particular, $\widehat{f \circ \gamma} = c_{I, \hat{x}_0}$, so $\hat{f}(z) = \widehat{f \circ \gamma}(1) = \hat{x}_0$ as required.

Now, it suffices to check that $\hat{f}$ is a continuous function. Let $U \subseteq \hat{X}$ be an open neighbourhood of $\hat{f}(z)$. We need to find an open neighbourhood $V \subseteq Z$ of z such that $\hat{f}(V) \subseteq U$.

First, we find a subset $U' \subset U$ with $\hat{f}(z) \in U'$ such that $p(U')$ is open and evenly covered. Since p is a covering map, there exists an open $W \subseteq X$ with $f(z) \in W$ and which is evenly covered. Hence $p^{-1}(W) = \coprod_{\alpha \in A} W_\alpha$, and $p(\hat{f}(z)) = f(z)$, so $\hat{f}(z) \in W_{\alpha_0}$ for some $\alpha_0 \in A$. Then, $W_{\alpha_0} \subseteq \hat{X}$ is an open set. Let $U' = U \cap W_{\alpha_0}$. Then $\hat{f}(z) \in U'$, and $p|_{W_{\alpha_0}} : W_{\alpha_0} \to W$ is a homeomorphism, so $p(U') = p_{\alpha_0}(U')$ is open and evenly covered.

Next, $f : Z \to X$ is continuous, so we need to find an open set $V' \subseteq Z$ with $z \in V'$ and $f(V') \subseteq p(U')$. Since Z is locally path-connected, there exists $V \subseteq V'$ which is an open path-connected set with $z \in V$.

Now we need to show V satisfies the continuity requirement, that $\hat{f}(V) \subseteq U$. Given $z' \in V$, let $\gamma' \in \Omega(V, z, z')$, which exists because V is path-connected. Then $\operatorname{Im} f \circ \gamma' \subseteq f(V) \subseteq p(U')$. Note that $\operatorname{Im} f \circ \gamma'$ is evenly covered. Hence $\tilde{\gamma}' = p_{\alpha_0}^{-1} \circ f \circ \gamma'$ is a lift of $f \circ \gamma'$ with $\tilde{\gamma}'(0) = p_{\alpha_0}^{-1}(f(z)) = \hat{f}(z)$. Then $\gamma\gamma' \in \Omega(Z, z_0, z')$, and $\widehat{f \circ (\gamma\gamma')} = \widehat{f \circ \gamma}\tilde{\gamma}'$ by the discussion at the beginning of the subsection. Hence $\hat{f}(z') = \widehat{f \circ (\gamma\gamma')}(1) = \tilde{\gamma}'(1) = p_{\alpha_0}^{-1} \circ f \circ \gamma'(1) \in U'$. So $\hat{f}(V) \subseteq U$ as required. □

4.4. Universal covers

Let $p : (\hat{X}, \hat{x}_0) \to (X, x_0)$ be a covering map. If $\gamma \in \Omega(X, x_0)$, let $\hat{\gamma} : I \to \hat{X}$ be its unique lift such that $\hat{\gamma}(0) = \hat{x}_0$, which exists by the path lifting property. Then there is a map $\varepsilon_p : \Omega(X, x_0) \to p^{-1}(x_0)$ by $\gamma \mapsto \hat{\gamma}(1)$, since $p(\hat{\gamma}(1)) = \gamma(1) = x_0$. By the corollary above, if $[\gamma_0] = [\gamma_1]$ in π_1, we have $\varepsilon_p(\gamma_0) = \varepsilon_p(\gamma_1)$. In particular, ε_p descends to a well-defined map from $\pi_1(X, x_0)$ to $p^{-1}(x_0)$.

Definition. A covering map $p : \hat{X} \to X$ is a *universal cover* if $\hat{X}$ is simply connected.

Example. $p : \mathbb{R} \to S^1$ defined by $x \mapsto e^{2\pi i x}$ is a universal cover of S^1, since $\mathbb{R}$ is contractible. $p_2 : \mathbb{R}^2 \to S^1 \times S^1 = T^2$ defined by $p_2(x, y) = (p(x), p(y))$ is a universal cover.

Proposition. If $p\colon (\hat{X}, \hat{x}_0) \to (X, x_0)$ is a universal cover, then $\varepsilon_p\colon \pi_1(X, x_0) \to p^{-1}(x_0)$ is a bijection of sets.

Proof. Suppose $\varepsilon_p[\gamma_0] = \hat{x}_1 = \varepsilon_p[\gamma_1]$. Then $\hat{\gamma}_0$ and $\hat{\gamma}_1$ are paths in $\Omega(\hat{X}, \hat{x}_0, \hat{x}_1)$. Since $\hat{X}$ is simply connected, $\hat{\gamma}_0 \sim_e \hat{\gamma}_1$. In particular, $\gamma_0 = p \circ \hat{\gamma}_0 \sim_e p \circ \hat{\gamma}_1 = \gamma_1$. Hence $[\gamma_0] = [\gamma_1]$, so ε_p is injective.

Given $\hat{x} \in p^{-1}(x_0)$, $\hat{X}$ is path-connected as it is simply connected, so there exists a path $\eta \in \Omega(\hat{X}, \hat{x}_0, \hat{x})$. Since $p(\hat{x}) = x_0$, we find $\gamma = p \circ \eta \in \Omega(X, x_0)$. Then $\eta = \hat{\gamma}$ is the unique lift of γ. In particular, $\varepsilon_p(\gamma) = \eta(1) = \hat{x}$, so ε_p is surjective. □

Example. Let $p\colon (\mathbb{R}, 0) \to (S^1, 1)$ be defined by $x \mapsto e^{2\pi i x}$. We have $p^{-1}(1) = \mathbb{Z}$. Then, $\varepsilon\colon \pi_1(S^1, 1) \to \mathbb{Z}$ is a bijection.

Theorem. $\varepsilon_p\colon \pi_1(S^1, 1) \to \mathbb{Z}$ is an isomorphism of groups.

Proof. It is a bijection, so it suffices to check that it is a homomorphism. Given $n \in \mathbb{Z}$, we can define $\varphi_n\colon \mathbb{R} \to \mathbb{R}$ by $\varphi_n(x) = x + n$. Then, $p \circ \varphi_n = p$. If $\gamma \in \Omega(S^1, 1)$, we can find a lift $\hat{\gamma}$ of γ with $\hat{\gamma}(0) = 0$. Then $p \circ \varphi_n \circ \hat{\gamma} = p \circ \hat{\gamma} = \gamma$, so $\varphi_n \circ \hat{\gamma}$ is a lift of γ with $\varphi_n \circ \hat{\gamma}(0) = n$.

Suppose $\varepsilon_p[\gamma] = n$, and $\varepsilon_p[\gamma'] = n'$. Then $\hat{\gamma}(1) = n$, $\hat{\gamma}'(1) = n'$, so $\varphi_n \circ \hat{\gamma}'$ is a lift of γ' that starts at n. Hence, $\widehat{\gamma\gamma'} = \hat{\gamma}(\varphi_n \circ \hat{\gamma}')$ is a lift of the composition of paths. Thus, $\varepsilon[\gamma\gamma'] = \widehat{\gamma\gamma'}(1) = \varphi_n(\hat{\gamma}'(1)) = n + n'$. So ε_p is a homomorphism. □

Corollary. S^1 is not contractible.

Example. Let $f\colon S^1 \to S^1$ be the identity map. Let $p\colon (\mathbb{R}, 0) \to (S^1, 1)$ be a covering map. Then there is no lift of f to $\mathbb{R}$. Otherwise, the identity map on $\mathbb{Z}$ would factor through the trivial group. This shows that the simply connected lifting property does not extend to all path-connected spaces.

4.5. Degree of maps on the circle

Lemma. Let $z \in S^1$, and $u, v \in \Omega(S^1, z, 1)$. Then, the isomorphisms $u_\sharp, v_\sharp\colon \pi_1(S^1, z) \to \pi_1(S^1, 1)$ are equal.

Proof. Consider $v_\sharp^{-1} \circ u_\sharp = (v^{-1})_\sharp \circ u_\sharp$. Note, $(v_\sharp^{-1} \circ u_\sharp)[\gamma] = [vu^{-1}\gamma u v^{-1}]$. Since $vu^{-1} \in \Omega(S^1, 1)$, we can write $[vu^{-1}\gamma u v^{-1}] = [\eta][\gamma][\eta^{-1}]$ where $\eta = vu^{-1}$. But this is exactly $[\gamma]$, since $\pi_1(S^1, 1) \simeq \mathbb{Z}$ is abelian. Hence $v_\sharp^{-1} \circ u_\sharp = \mathrm{id}$, and by symmetry, $u_\sharp^{-1} \circ v_\sharp = \mathrm{id}$. □

Definition. Let $f\colon S^1 \to S^1$, $f(1) = z$. Then choose $u \in \Omega(S^1, z, 1)$, then $f_\star\colon \pi_1(S^1, 1) \to \pi_1(S^1, z)$, giving $u_\sharp \circ f_\star\colon \pi_1(S^1, 1) \to \pi_1(S^1, 1)$. This is a homomorphism $\mathbb{Z} \to \mathbb{Z}$, so is uniquely determined by its action on 1. We define the *degree* of f, written $\deg f$, to be $(u_\sharp \circ f_\star)(1)$.

By the above lemma, this definition does not depend on the choice of path u.

Example. Let $\gamma_n \in \Omega(S^1, 1)$ be given by $\gamma_n(t) = e^{2\pi i n t}$ for $n \in \mathbb{Z}$. Then $\hat{\gamma}_n(t) = nt$, so $\varepsilon_p[\gamma_n] = n$. The integers n correspond to the classes $[\gamma_n]$ in $\pi_1(S^1, 1)$.

Let $f_n = \overline{\gamma}_n : S^1 \to S^1$, so $f_n(z) = z^n$. Then $f_n \circ \gamma_1 = \gamma_n$, so $f_{n\star}[\gamma_1] = [\gamma_n]$. Hence the degree of f_n is n.

Proposition. The degree of $f_n : S^1 \to S^1$, defined by $z \mapsto z^n$, is n. If $g_0, g_1 : S^1 \to S^1$, then $g_0 \sim g_1$ if and only if $\deg g_0 = \deg g_1$. $g : S^1 \to S^1$ extends to $G : D^2 \to S^1$ if and only if $\deg g = 0$.

Proof. Suppose $g_0 \sim g_1$ via $H : S^1 \times I \to S^1$. Let $u(t) = H(1, t)$, so $g_{1\star} = u_\sharp \circ g_{0\star}$, where $u \in \Omega(S^1, g_0(1), g_1(1))$. Let $v \in \Omega(S^1, g_1(1), 1)$. Then $uv \in \Omega(S^1, g_0(1), 1)$, and so $\deg g_1 = v_\sharp \circ g_{1\star}(1) = v_\sharp(u_\sharp \circ g_0(1)) = (uv)_\sharp \circ g_{0\star}(1) = \deg g_0$, since $u_\sharp[\gamma] = [u^{-1}\gamma u]$ so $(u \circ v)_\sharp = v_\sharp \circ u_\sharp$.

Conversely, it suffices to show that $g \sim f_{\deg g}$ by transitivity. Suppose $g(1) = 1$. Then $g = \overline{\gamma}$ where $\gamma = g \circ \gamma_1$. Then $\deg g = g_\star(1) = [g \circ \gamma_1] = [\gamma] \in \pi_1(S^1, 1)$. In particular, if $\deg g = n$, we have $\gamma \sim \gamma_n$, so $g = \overline{\gamma} \sim \overline{\gamma}_n = f_n$.

In general, if $g(1) = e^{2\pi i x}$, then $g \sim g_0$ where $g_0(z) = e^{-2\pi i x} g(z)$ via $g_t(z) = e^{-2\pi i t x} g(z)$. Then $g \sim g_0$ so $\deg g = \deg g_0$, so in particular $g \sim g_0 \sim \gamma_{\deg g}$.

g extends to D^2 if and only if $g \sim c_{S^1, z_0}$ for some $z_0 \in S^1$. Equivalently, $g \sim c_{S^1,1} = f_0$, so $\deg g = 0$ by above. □

4.6. Fundamental theorem of algebra

Let $p : \mathbb{C} \to \mathbb{C}$ be a polynomial, so $p(w) = w^n + a_{n-1}w^{n-1} + \dots + a_0 = w^n + q(w)$.

Lemma. Let $R_0 = \max\left\{1, \sum_{i=0}^{n-1} |a_i|\right\}$. Then if $|w| > R_0$, $|w^n| > |q(w)|$.

Proof. Consider

$$\frac{|q(w)|}{|w^{n-1}|} \leq \sum_{i=0}^{n-1} |a_i| |w|^{i-n+1}$$

Hence, if $|w| > 1$, each term $|w|^{i-n+1}$ is at most one.

$$\sum_{i=0}^{n-1} |a_i| |w|^{i-n+1} \leq \sum_{i=0}^{n-1} |a_i| \leq R_0$$

Hence $\frac{|q(w)|}{|w|^n} < \frac{R_0}{|w|} < 1$. □

Consider $g_0, g_1 : S^1 \to \mathbb{C} \setminus \{0\}$ given by $g_0(z) = (Rz)^n$ for some fixed $R > R_0$, and $g_1(z) = p(Rz)$. Then $g_0 \sim g_1$ via $g_t(z) = p_t(Rz)$ where $p_t(w) = w^n + tq(w)$. This map has codomain $\mathbb{C} \setminus \{0\}$ by the above lemma. Let $\pi : \mathbb{C} \setminus \{0\} \to S^1$ be the radial projection $w \mapsto \frac{w}{|w|}$. Then $\pi \circ g_0, \pi \circ g_1 : S^1 \to S^1$ are homotopic maps. Therefore, $n = \deg(\pi \circ g_0) = \deg(\pi \circ g_1)$.

Theorem. If $n > 0$, p has a root $w_0 \in \mathbb{C}$.

Proof. If $p(w) \neq 0$ for all w, $p : \mathbb{C} \to \mathbb{C} \setminus \{0\}$, so g_1 extends to $G_1 : D^2 \to \mathbb{C} \setminus \{0\}$ given by $G_1(z) = p(Rz)$. Then $\pi \circ G_1$ is an extension of $\pi \circ g_1$. So $n = \deg \pi \circ g_1 = 0$, so we have a constant polynomial. □

4.7. Wedge product

Definition. Let (X_i, x_i) be pointed spaces. The *wedge product* $\bigvee_{i=1}^n (X_i, x_i) = {\coprod_{i=1}^n (X_i, x_i)}/{\sim}$ for the equivalence relation $\sim$ generated by $x_i \sim x_j$. For $n = 2$, we also write $(X_1, x_1) \vee (X_2, x_2)$ for $\bigvee_{i=1}^2 (X_i, x_i)$.

If each X_i has the property that for any $x_i, x_i' \in X_i$, there exists a homeomorphism $\varphi : X_i \to X_i$ such that $\varphi(x_i) = \varphi(x_i')$, then the particular choice of base point used in the wedge product does not matter, and the expression $\bigvee_{i=1}^n X_i = \bigvee_{i=1}^n (X_i, x_i)$ is well-defined up to homeomorphism independent of the choice of the x_i.

Example. Consider the figure-eight $S^1 \vee S^1$. There are inclusion maps $\iota_1, \iota_2 : (S^1, 1) \to (S^1 \vee S^1, x_0)$ where x_0 is the point at which the two circles are joined. Let $a = \iota_{1\star}(1) \in \pi_1(S^1 \vee S^1, x_0)$, and similarly let $b = \iota_{2\star}(1) \in \pi_1(S^1 \vee S^1, x_0)$. The universal cover of $S^1 \vee S^1$ is the infinite regular 4-valent tree, $T_\infty(4)$. If $T_n(4)$ is the regular 4-valent tree of depth n, $T_\infty(4) = \bigcup_{n=1}^\infty T_n(4)$, so $U \subseteq T_\infty(4)$ is open if and only if $U \cap T_n(4)$ is open for all n. There is a covering map from $T_\infty(4)$ to $S^1 \vee S^1$ by mapping each edge to one of the circles. $T_\infty(4)$ is simply connected, because the interval I is compact, so if $\gamma : I \to T_\infty(4)$, $\operatorname{Im} \gamma \subseteq T_n(4)$ for some n, and each of the finite trees is contractible and therefore simply connected.

In particular, there is a bijection $\pi_1(S^1 \vee S^1, x_0) \to p^{-1}(\{x_0\})$ given by $[\gamma] \to \varepsilon_p(\gamma)$. Here, $\varepsilon_p(ab) = \widehat{ab}(1)$, but $\varepsilon_p(ba) = \widehat{ba}(1) \neq \widehat{ab}(1)$. In $\pi_1(S^1 \vee S^1, x_0)$, $ab \neq ba$, so $\pi_1(S^1 \vee S^1, x_0)$ is not abelian.

4.8. Covering transformations

Definition. Let $p_i : \hat{X}_i \to X$ be covering maps for $i = 1, 2$. A *covering transformation* $p : (p_1, \hat{X}_1) \to (p_2, \hat{X}_2)$ is a map $p : \hat{X}_1 \to \hat{X}_2$ such that $p_2 \circ p = p_1$.

$$\begin{tikzcd}
\hat{X}_1 \arrow[rr, dashed, "p"] \arrow[dr, "p_1"'] & & \hat{X}_2 \arrow[dl, "p_2"] \\
& X &
\end{tikzcd}$$

Remark. We can think of p as a lift of p_1 to $\hat{X}_2$.

$$\begin{tikzcd}
 & \hat{X}_2 \arrow[d, "p_2"] \\
\hat{X}_1 \arrow[ur, dashed, "p"] \arrow[r, "p_1"'] & X
\end{tikzcd}$$

Example. Let $p_1 : S^1 \to S^1$ be defined by $z \mapsto z^6$, and $p_2 : S^1 \to S^1$ be defined by $z \mapsto z^2$. Then $p : (p_1, S^1) \to (p_2, S^1)$ defined by $z \mapsto z^3$ is a covering transformation.

$$\begin{tikzcd}
S^1 \arrow[rr, dashed, "z\mapsto z^3"] \arrow[dr, "z\mapsto z^6"'] & & S^1 \arrow[dl, "z\mapsto z^2"] \\
 & S^1 &
\end{tikzcd}$$

Lemma. Let X be locally path-connected. If $p : (p_1, \hat{X}_1) \to (p_2, \hat{X}_2)$ is a covering transformation, $p : \hat{X}_1 \to \hat{X}_2$ is a covering map.

$$\begin{tikzcd}
\hat{X}_1 \arrow[d, dashed, "p"] \arrow[dd, bend right, "p_1"'] \\
\hat{X}_2 \arrow[d, "p_2"] \\
X
\end{tikzcd}$$

Proof. Given $x_2 \in \hat{X}_2$, we find an open evenly covered neighbourhood U_{x_2}. Let $x = p_2(x_2) \in X$. Then p_1, p_2 are covering maps of X, so there exist open neighbourhoods U_1, U_2 of x such that U_i is evenly covered by p_i. Then $U = U_1 \cap U_2$ is open and evenly covered by p_1 and p_2. Since X is locally path-connected, let $V \subseteq U$ be an open neighbourhood of x that is path-connected. Then $p_1^{-1}(V) = \coprod_{\alpha\in A} V_\alpha$ and $p_2^{-1}(V) = \coprod_{\beta\in B} V_\beta$, where $V_\alpha \simeq V \simeq V_\beta$ are all path-connected. Let $x_\alpha = p_{1,\alpha}^{-1}(x)$, and $x_\beta = p_{2,\beta}^{-1}(x)$. Then $p_2(p(x_\alpha)) = p_1(x_\alpha) = x$, so $p(x_\alpha) = x_\beta$ for some $\beta \in B$. Now, V_α, V_β are path-connected, so $p(V_\alpha) \subseteq V_\beta$ since each V_β is a (maximal) path-connected component of $p_2^{-1}(V)$. Therefore, $p|_{V_\alpha} : V_\alpha \to V_\beta$ satisfies $p_{2,\beta} \circ p|_{V_\alpha} = p_{1,\alpha}$, so $p|_{V_\alpha} = p_{2,\beta}^{-1} \circ p_{1,\alpha}$ is a homeomorphism. In particular, $p^{-1}(V_\beta) = \coprod_{\alpha\in V, p(x_\alpha)=x_\beta} V_\alpha$, and $p|_{V_\alpha} : V_\alpha \to V_\beta$ is a homeomorphism. So V_β is evenly covered, so p is indeed a covering map. □

4.9. Uniqueness of universal covers

Let X be a locally path-connected space, and $q : (\widetilde{X}, \tilde{x}_0) \to (X, x_0)$ be a universal cover. Let $p : (\hat{X}, \hat{x}_0) \to (X, x_0)$.

Lemma. If $p : \hat{Y} \to Y$ is a bijective covering map, then p is a homeomorphism.

Proof. p is continuous and bijective, therefore $p^{-1} : Y \to \hat{Y}$ exists as a map of sets. We must show that this map is continuous. Since p is a covering map, Y has an open cover $\{U_y \mid y \in Y\}$ such that U_y is evenly covered. In particular, $p^{-1}|_{U_y} : U_y \to p^{-1}(U_y)$ is a homeomorphism. Hence p^{-1} is continuous. □

Recall that if $p_i : \hat{X}_i \to X$ are covering maps, a covering transformation from $(p_1, \hat{X}_1)$ to $(p_2, \hat{X}_2)$ is a lift $\hat{p}_1$ of p_1 to X_2. $\hat{p}_1$ is a covering isomorphism if it is bijective. Then, by the lemma, it is a homeomorphism.

Proposition. Let X be a locally path-connected space, and $q\colon (\widetilde{X}, \tilde{x}_0) \to (X, x_0)$ be a universal cover. Let $p\colon (\hat{X}, \hat{x}_0) \to (X, x_0)$. Then there is a unique covering transformation $\hat{q}\colon (p, \hat{X}) \to (q, \widetilde{X})$

$$\begin{tikzcd}
 & (\hat{X}, \hat{x}_0) \arrow[d, "p"] \\
(\widetilde{X}, \tilde{x}_0) \arrow[ur, dashed, "\hat{q}"] \arrow[r, "q"'] & (X, x_0)
\end{tikzcd}$$

Proof. Note that $\widetilde{X}$ is simply connected, and since X is locally path-connected, so is $\widetilde{X}$. So existence and uniqueness of $\hat{q}$ is exactly the simply connected lifting property. □

Corollary. If p is also a universal cover, $\hat{q}$ is a covering isomorphism, and in particular, $\hat{X} \simeq \widetilde{X}$.

Proof. $\widetilde{X}$ is simply connected, so $\hat{q} : \widetilde{X} \to \hat{X}$ is a universal cover. Hence, there is a bijection between points $\hat{q}^{-1}(\hat{x})$ and elements $\pi_1(\hat{X}, \hat{x})$. But this is the one-element set, since $\hat{X}$ is simply connected. So $\hat{q}^{-1}(\hat{x})$ has a single element, and so $\hat{q}$ is a bijection. □

Equivalently, if $q : (\widetilde{X}, \tilde{x}_0) \to (X, x_0)$ and $q' : (\widetilde{X}', \tilde{x}_0') \to (X, x_0)$ are universal covers, there is a unique covering isomorphism $\hat{q} : (\widetilde{X}, \tilde{x}_0) \to (\widetilde{X}', \tilde{x}_0')$.

4.10. Deck groups

Definition. The *deck group* $G_D(p)$ is the set of covering automorphisms $g\colon (p, \hat{X}) \to (p, \hat{X})$, which forms a group under composition $gf = g \circ f$. This has a left action on $\hat{X}$ by $g \cdot \hat{x} = g(\hat{x})$.

Example. Let $p\colon (\mathbb{R}, 0) \to (S^1, 1)$. The deck group $G_D(p)$ is exactly

$$\{g_n : \mathbb{R} \to \mathbb{R} \mid g_n(t) = t + n\} \simeq \mathbb{Z}$$

In this case, $G_D(p) \simeq \pi_1(S^1, 1)$.

Example. There is a bijection between $G_D(q)$ and $q^{-1}(x_0)$, by $g \mapsto g(\tilde{x}_0)$, by the above proposition with $\hat{X} = \widetilde{X}$.

Theorem. Let $q\colon (\widetilde{X}, \tilde{x}_0) \to (X, x_0)$ be a universal cover. Then $G_D(q) \simeq \pi_1(X, x_0)$.

Proof. There is a bijection between $\pi_1(X, x_0)$ and $q^{-1}(x_0)$ since q is a universal cover. By the above example, $q^{-1}(x_0)$ is in bijection with $G_D(q)$. In particular, we can map $[\gamma] \in \pi_1(X, x_0)$ to $\tilde{\gamma}(1) \in q^{-1}(x_0)$, where $\tilde{\gamma}$ is the unique lift of γ starting at $\tilde{x}_0$, and $g(\tilde{x}_0) \in q^{-1}(x_0)$ is mapped to $g \in G_D(q)$. We need to check that this composed map is a homomorphism: it is already a bijection of sets.

$[\gamma\gamma']$ is mapped to $\widetilde{\gamma\gamma'}(1) = \tilde{\gamma}(g_{\tilde{\gamma}(1)} \circ \tilde{\gamma})$ where $g_{\tilde{\gamma}(1)}$ is the unique element of $G_D(q)$ with $g_{\tilde{\gamma}(1)}(x_0) = \tilde{\gamma}(1)$. Since $g_{\tilde{\gamma}(1)} \circ \tilde{\gamma}'$ is a lift of $\tilde{\gamma}'$ starting at $\tilde{\gamma}(1)$, we have $\widetilde{\gamma\gamma'}(1) = (g_{\tilde{\gamma}(1)} \circ \tilde{\gamma}')(1) = g_{\tilde{\gamma}(1)}(\tilde{\gamma}'(1)) = g_{\tilde{\gamma}(1)}(g_{\tilde{\gamma}'(1)}(\tilde{x}_0))$. So $\widetilde{\gamma\gamma'}(1)$ is the image of $\tilde{x}_0$ under $g_{\tilde{\gamma}(1)} \circ g_{\tilde{\gamma}'(1)}$, so this is indeed a homomorphism. □

4.11. Correspondence of subgroups and covers

Proposition. Let $G = G_D(q) \simeq \pi_1(X, x_0)$. If $H \leq G$ is a subgroup, we have a tower of covering maps

$$\begin{tikzcd}
\widetilde{X} \arrow[d, dashed, "\pi_H"] & 1 \\
X_H \arrow[d, "p_H"] & H \arrow[u] \\
X & G \arrow[u]
\end{tikzcd}$$

where $X_H = H \setminus \widetilde{X}$ is the quotient given by $h \cdot x \sim x$ for all $h \in H$. In particular, $\pi_H : \widetilde{X} \to H \setminus \widetilde{X}$ is the quotient map, and $p_H : X_H \to X$ is given by $p_H(H \cdot x) = q(x)$. This is well-defined because $q \circ h = q$ as h is a deck transformation. In particular, if $H = G$, p_G is a covering isomorphism, so $X \simeq G \setminus \widetilde{X}$.

A universal covering map is a quotient by the action of $G_D(q) \simeq \pi_1(X, x_0)$.

Proof. Let $x \in X$. Then choose U_x to be evenly covered by q. Then $q^{-1}(U_x) = \coprod_{\alpha \in A} U_\alpha = \coprod_{g \in G_D(q)} g \cdot U_{\alpha_0}$ for $\tilde{x}_0 \in U_{\alpha_0}$. Then $p_H^{-1}(U_x) = \coprod_{\beta = gH \in \text{cosets of } H} U_\beta$. Then $\pi_H^{-1}(U_\beta) = \coprod_{gh \in gH} gh \cdot U_{\alpha_0}$, and $p_H^{-1}(U_x) = \coprod U_\beta$. So each is evenly covered. □

Definition. $p : \widetilde{X} \to X$ is a *normal cover* if $G_D(p)$ acts transitively on $p^{-1}(x_0)$.

Example. The universal cover q is always a normal cover.

Proposition. Let $p : (\hat{X}, \hat{x}_0) \to (X, x_0)$ be a covering map. Then $p_\star : \pi_1(\hat{X}, \hat{x}_0) \to \pi_1(X, x_0)$ is injective. In particular, $\operatorname{Im} p_\star \simeq \pi_1(\hat{X}, \hat{x}_0)$ is a subgroup of $\pi_1(X, x_0)$.

Proof. If $p_\star[\gamma_0] = p_\star[\gamma_1]$, we have $p \circ \gamma_0 \sim_e p \circ \gamma_1$, so $p \circ \hat{\gamma}_0 \sim_e p \circ \hat{\gamma}_1$, so $\gamma_0 \sim_e \gamma_1$. In particular, $[\gamma_0] = [\gamma_1]$. □

Let $q\colon (\widetilde{X}, \tilde{x}_0) \to (X, x_0)$ be a universal cover, so $\widetilde{X}$ and hence X are path-connected. Suppose further that X is locally path-connected, so $\widetilde{X}$ is also locally path-connected. Consider

$$\begin{aligned} S(X,x_0) &= \{H \leq \pi_1(X,x_0)\} \\ C(X,x_0) &= {\big\{(p,\hat{X},\hat{x}_0) \mid p\colon (\hat{X},\hat{x}_0) \to (X,x_0) \text{ is a covering map}, \hat{X} \text{ is path-connected}\big\}}\big/{\sim} \end{aligned}$$

where $(p,\hat{X},\hat{x}_0) \sim (p',\hat{X}',\hat{x}_0')$ if there is a covering isomorphism $q\colon (p,\hat{X}) \to (p',\hat{X}')$ mapping $\hat{x}_0 \mapsto \hat{x}_0'$. Let $\alpha\colon S(X,x_0) \to C(X,x_0)$ be given by $\alpha(H) = (p_H, X_H, x_{0,H})$, where $X_H = H \setminus \widetilde{X}$, so $\widetilde{X} \xrightarrow{\pi_H} X_H \xrightarrow{p_H} X$ mapping $\tilde{x}_0$ to $x_{0,H}$. Let $\beta\colon C(X,x_0) \to S(X,x_0)$ be defined by $(p,\hat{X},\hat{x}_0) \mapsto p_\star(\pi_1(\hat{X},\hat{x}_0))$.

Theorem. α, β are inverses, and hence bijections.

Remark. The entire group $G = \pi_1(X,x_0)$ is mapped to (id, X, x_0). The trivial group $1 \subseteq G$ is mapped to the universal cover $(q, \widetilde{X}, \tilde{x}_0)$. The index $[G : H]$ is exactly $|p_H^{-1}(x_0)|$. A conjugation $g^{-1}Hg$ corresponds to a change of base point $(p_H, X_H, \hat{\gamma}(1))$, where $g = [\gamma]$ and $\hat{\gamma}\colon I \to X_H$ is a lift of γ with $\hat{\gamma}(0) = x_{0,H}$. If $H \trianglelefteq G$ is a normal subgroup, p_H is a normal covering. The quotient ${G}\big/{H}$ corresponds to the deck group $G_D(p_H)$.

Proof. Consider $\beta(\alpha(H)) = p_{H\star}(\pi_1(X_H, x_{0,H}))$. There are isomorphisms

$$H \to \pi_1(X_H, x_{0,H}) \to p_{H\star}(\pi_1(X,x_0))$$

mapping

$$[\gamma] \mapsto [\pi_H \circ \tilde{\gamma}] \mapsto [p_H \circ \pi_H \circ \tilde{\gamma}] = [\pi_G \circ \tilde{\gamma}] = [\gamma]$$

where $\tilde{\gamma}$ is a lift of γ such that $\tilde{\gamma}(0) = \tilde{x}_0$. Hence $\beta(\alpha(H)) = H$.

Conversely, consider $\alpha(\beta((p,\hat{X},\hat{x}_0))) = (p_H, X_H, x_{0,H})$ where $H = p_\star(\pi_1(X,x_0))$. Consider

$$\begin{tikzcd}
(X_H, x_{0,H}) \arrow[r, "p'"] & (\hat{X}, \hat{x}_0) \arrow[d, "p"] \\
(\widetilde{X}, \tilde{x}_0) \arrow[u, "\pi_H"] \arrow[ur, "\hat{q}"] \arrow[r, "q"'] & (X, x_0)
\end{tikzcd}$$

We claim that $\hat{q} = p' \circ \pi_H$, where p' is a covering isomorphism. If we can show this, we have $(p_H, X_H, x_{0,H}) \sim (p, \hat{X}, \hat{x}_0)$, so $\alpha \circ \beta$ is the identity on $C(X,x_0)$. If $h \in H = p_\star(\pi_1(\hat{X},\hat{x}_0))$, $h = [p \circ \gamma]$ for some $\gamma \in \Omega(\hat{X},\hat{x}_0)$. Then $\hat{q}(\tilde{x}) = \widehat{q \circ \eta_{\tilde{x}}}(1)$ where $\eta_{\tilde{x}} \in \Omega(\widetilde{X}, \tilde{x}_0, \tilde{x})$. Then $\eta_{h\cdot\tilde{x}} = \eta_{h\circ\tilde{x}_0}(h \circ \eta_{\tilde{x}})$, so $q \circ \eta_{h\cdot\tilde{x}} = (q \circ \eta_{h\cdot\tilde{x}_0})(q \circ \eta_{\tilde{x}}) = (p \circ \gamma)(q \circ \eta_{\tilde{x}})$, so in particular, $\widehat{q \circ \eta_{h\cdot\tilde{x}}} = (\gamma)(\widehat{q \circ \eta_{\tilde{x}}})$. Hence $\hat{q}(h \cdot \tilde{x}) = (q \circ \eta_{h\cdot\tilde{x}})(1) = \widehat{q \circ \eta_{\tilde{x}}}(1) = \hat{q}(\tilde{x})$, so $\hat{q}$ factors as shown. $\hat{X}$ is connected, so p' is surjective, so it is bijective and hence a covering isomorphism. □

5. Seifert–Van Kampen theorem

5.1. Free groups and presentations

Consider $\pi_1(S^1 \vee S^1, x_0)$ where x_0 is the wedge point. The universal cover is the infinite 4-valent tree $T_\infty(4)$, so $\pi_1(S^1 \vee S^1)$ is in bijection with $q^{-1}(x_0)$, the vertices of $T_\infty(4)$. Let $\tilde{x}_0$ be one such vertex. If $\tilde{x}$ is a vertex, there is a unique shortest path from $\tilde{x}_0$ to $\tilde{x}$. This gives an 'address' for $\tilde{x}$ in $T_\infty(4)$ given by recording the type and direction of each edge used in the path. The set of such 'addresses' is in bijection with the set of *reduced words* $w = \ell_1 \dots \ell_r$ where $r \in \mathbb{N}$, and each l_i is one of a, a^{-1}, b, b^{-1}, such that w does not contain any substring of the form $aa^{-1}, a^{-1}a, bb^{-1}b^{-1}b$. Then each word w corresponds to an element $w \in \pi_1(S^1 \vee S^1, x_0)$, the image of the shortest path under q. Note that the multiplication ww' in $\pi_1(S^1 \vee S^1, x_0)$ corresponds to concatenation of words ww' and then the reduction of substrings such as aa^{-1}.

Definition. A *free group* with generating set S is a group F_S and a subset $S \subseteq F_s$ such that if G is a group and $\varphi : S \to G$ is a map of sets, there is a unique homomorphism $\Phi : F_s \to G$ with $\Phi|_S = \varphi$.

$$\begin{tikzcd}
 & F_S \arrow[d, dashed, "\Phi"] \\
S \arrow[ur] \arrow[r, "\varphi"'] & G
\end{tikzcd}$$

Remark. The action of taking the free group of a set is a functor from **Set** to **Grp**, and it is left adjoint to the forgetful functor from **Grp** to **Set**. This property is known as the universal property of the free group.

Example. $\pi_1(S^1 \vee S^1) \simeq F_{\{a,b\}}$. Indeed, given $\varphi : \{a,b\} \to G$, we define $\Phi(\ell_1 \dots \ell_r) = \varphi(\ell_1) \dots \varphi(\ell_r)$, where we extend φ to all of $\{a, a^{-1}, b, b^{-1}\}$ by defining $\varphi(a^{-1}) = \varphi(a)^{-1}$ and $\varphi(b^{-1}) = \varphi(b)^{-1}$. This is a homomorphism: indeed,

$$\Phi(ww') = \varphi(\ell_1) \dots \varphi(\ell_k)\varphi(\ell_1') \dots \varphi(\ell_k') = \Phi(w)\Phi(w')$$

cancelling substrings of the form aa^{-1} as required. The homomorphism is unique as required for the universal property of the free group.

Lemma. Let F_S, F_T be free groups on sets $S \subseteq F_S, T \subseteq F_T$. Let $\varphi : S \to T$ be a bijection. Then $\Phi : F_S \to F_T$ is an isomorphism.

Proof. Let $\psi = \varphi^{-1}$. Since F_T is free, there exists a homomorphism $\Psi : F_T \to F_S$ such that $\Psi|_T = \psi$. Then $\Psi \circ \Phi : F_S \to F_S$ has the property that for all $s \in S$, we have $\psi \circ \varphi(s) = s$. F_S is free, so there is a unique homomorphism $\alpha : F_S \to F_S$ mapping $s \in S$ to s. So $\alpha = \mathrm{id}_{F_S}$. Hence $\Psi \circ \Phi = \mathrm{id}_{F_S}$, so by symmetry, they are inverse functions. □

Corollary. If F_S, F_S' are free groups generated by S, $F_S \simeq F_S'$. So the isomorphism type of F_S depends only on $|S|$, the cardinality of S.

We therefore can write F_n for *the* free group (up to isomorphism) generated by n elements $a_1, \dots, a_n$. Let $X = \bigvee_{i=1}^n S^1$ where x_0 is the wedge point, with inclusion maps $j_n : S^1 \to X$. Let $a_i = j_{i\star}(1)$ for $1 \in \pi_1(S^1, 1)$ be a generator. Then X has universal cover $\widetilde{X} = T_\infty(2n)$, the infinite regular $2n$-valent tree. In particular, $\pi_1(X, x_0)$ is the set of reduced words in $\{a_1^{\pm 1}, \dots, a_n^{\pm 1}\}$, which is isomorphic to F_{2n}.

5.2. Presentations

Definition. Let G be a group and $S \subseteq G$ be a subset. Let $\mathcal{S}_S = \{H \leq G \mid S \subseteq H\}$, then let $\langle S \rangle = \bigcap_{H \in \mathcal{S}_S} H$ be the smallest subgroup of G containing S, known as the *subgroup generated by* S. Similarly, let $\mathcal{N}_S = \{N \trianglelefteq G \mid S \subseteq H\}$, and let $\langle\!\langle S \rangle\!\rangle = \bigcap_{H \in \mathcal{N}_S} H$ be the smallest normal subgroup of G containing S, called the *subgroup normally generated by* S.

Note that $\langle S \rangle$ is nonempty since $1 \in H$ for all $H \in \mathcal{S}_S$.

If $\langle S \rangle = G$, we say that S *generates* G. If so, there is a unique homomorphism $\Phi_S : F_S \to G$ that maps s to s. $\operatorname{Im} \Phi_S \leq G$, and it contains S, so Φ_S is surjective.

Definition. Given a set S and $R \subseteq F_S$, we define $\langle S \mid R \rangle = {}^{F_S}\!/\!_{\langle\!\langle R \rangle\!\rangle}$. If in addition $\langle\!\langle R \rangle\!\rangle = \ker \Phi_S$, then $G \simeq {}^{F_S}\!/\!_{\ker \Phi_S} = {}^{F_S}\!/\!_{\langle\!\langle R \rangle\!\rangle}$. We say $\langle S \mid R \rangle$ is a *presentation* for G.

Proposition. Any group G admits a presentation.

Proof. Clearly $\langle G \rangle = G$, so let $S = G$. Let $R = \ker \Phi_G$, where $\Phi_G : F_G \to G$. Then by construction, ${}^{F_S}\!/\!_{\langle\!\langle R \rangle\!\rangle} = {}^{F_S}\!/\!_{\ker \Phi_G} \simeq G$. □

Remark. These presentations are very large. It is often more useful to consider *finite* presentations of G, where both S and R are finite.

Example. $\langle a, b \mid \rangle \simeq F_2$. $\langle a \mid \rangle \simeq F_1 = \pi_1(S^1, 1) \simeq \mathbb{Z}$. $\langle a \mid a^3 \rangle \simeq {}^{\mathbb{Z}}\!/\!_{3\mathbb{Z}}$. $\langle a, b \mid ab^{-3} \rangle \simeq \mathbb{Z}$.

Proposition. Let $\langle S \mid R \rangle$ be a presentation, let $a \notin S$, and let $w \in F_S$. Then $\langle S \mid R \rangle \simeq \langle S \cup \{a\} \mid R \cup \{aw^{-1}\} \rangle$.

Proof. We have homomorphisms $\varphi : \langle S \mid R \rangle \to \langle S \cup \{a\} \mid R \cup \{aw^{-1}\} \rangle$ mapping $s \in S$ to s, and $\psi : \langle S \cup \{a\} \mid R \cup \{aw^{-1}\} \rangle \to \langle S \mid R \rangle$ mapping $s \in S$ to s and a to w. These are inverses. □

There are other operations we can apply to presentations. If $w \in R$, we can replace w with a conjugate sws^{-1} for $s \in S$, and it leaves the group unchanged. For example, $\langle ab \mid abb \rangle = \langle ab \mid bab \rangle$. Also, if $w_1, w_2 \in R$, we can replace w_1 with $w_1 w_2$, so for example,

$$\langle ab \mid babb, abb \rangle = \langle ab \mid b, abb \rangle \simeq \langle a \mid a \rangle \simeq 1$$

Theorem. Given a finite set S and a finite set of relations $R \subseteq F_S$, there is no algorithm to determine if $\langle S \mid R \rangle \simeq 1$.

5.3. Covering with a pair of open sets

Theorem. Let $U_1, U_2 \subseteq X$ be open, and $U_1 \cap U_2$ be path-connected with $x_0 \in U_1 \cap U_2$ and $U_1 \cup U_2 = X$. Then $\iota_{1\star}(\pi_1(U_1, x_0)) \cup \iota_{2\star}(\pi_1(U_2, x_0))$ generates $\pi_1(X, x_0)$, where $\iota_i : U_i \to X$ is the inclusion.

Proof. $\{U_1, U_2\}$ is an open cover of X, so if $\gamma \in \Omega(X, x_0)$, we have $\{\gamma^{-1}(U_1), \gamma^{-1}(U_2)\}$ is an open cover of I. By the Lebesgue covering lemma, we can find $n \in \mathbb{N}$ such that $\left[\frac{j}{n}, \frac{j+1}{n}\right]$ lies entirely inside $\gamma^{-1}(U_1)$ or $\gamma^{-1}(U_2)$ for all j. Each interval $\left[\frac{j}{n}, \frac{j+1}{n}\right]$ with the label 1 or 2 accordingly; if it lies in both, choose an arbitrary label. Let $0 = t_0 < t_1 < \dots < t_k = 1$ be the points of the form $\frac{j}{n}$ where the labelling changes. Let $I_i = [t_{i-1}, t_i]$ for each $i \in \{0, \dots, k\}$. Let $\gamma_i = \gamma|_{I_i}$, so $\gamma(t_i) \in U_1 \cap U_2$, and $\gamma(I_i) \subseteq U_{i \bmod 2}$ without loss of generality. Note that we can write γ as the composition of paths $\gamma = \gamma_1 \dots \gamma_k$.

Let $\eta_1, \dots, \eta_{k-1}$ be paths with $\eta_i \in \Omega(U_1 \cap U_2, \gamma(t_i), x_0)$, which exists since $U_1 \cap U_2$ is path-connected. Then

$$\gamma \sim_e \gamma_1\eta_1\eta_1^{-1}\gamma_2\eta_2\eta_2^{-1} \dots \eta_{k-1}\eta_{k-1}^{-1}\gamma_k = \underbrace{(\gamma_1\eta_1)}_{\delta_1}\underbrace{(\eta_1^{-1}\gamma_2\eta_2)}_{\delta_2}\eta_2^{-1} \dots \eta_{k-1}\underbrace{(\eta_{k-1}^{-1}\gamma_k)}_{\delta_k}$$

Then each $\delta_i \in \Omega_i(U_1, x_0)$, so $[\delta_i] \in \operatorname{Im} \iota_{(i \bmod 2)\star}$. So $[\gamma] = [\delta_1][\delta_2] \dots [\delta_k]$ is a product of elements in $\iota_{1\star}(\pi_1(U_1, x_0)) \cup \iota_{2\star}(\pi_1(U_2, x_0))$, so $[\gamma]$ lies in the subgroup they generate. □

Corollary. Let $U_1, U_2 \subseteq X$ be open and simply connected with $U_1 \cup U_2 = X$, and let $U_1 \cap U_2$ be path-connected and contain x_0. Then X is simply connected.

Proof. $\pi_1(X, x_0)$ is generated by $\iota_{1\star}(\pi_1(U_1, x_0)) \cup \iota_{2\star}(\pi_1(U_2, x_0)) = \{1\}$. □

Example. $S^n = U^+ \cup U^-$, where $U^+ = S^n = \{(1, 0, \dots, 0)\}$ and $U^- = S^n - \{(-1, 0, \dots, 0)\}$. Then $U^+ \simeq U^- \simeq \mathbb{R}^n$ by stereographic projection. $U^+ \cap U^- \simeq \mathbb{R}^n - \{0\}$. Hence $\pi_1(U^\pm, x_0) = 1$ since $\mathbb{R}^n$ is contractible. $U^+ \cap U^-$ is path-connected if $n > 1$, so $\pi_1(S^n, x_0) = 1$ for $n > 1$.

Example (attaching a disk). If $f : S^1 \to X$ with $f(1) = x_0$, let $X \cup_f D^2 = {}^{X \amalg D^2}\!/_{\sim}$, where $\sim$ is the smallest equivalence relation such that $z \sim f(z)$ for $z \in S^1$. Let π be the quotient map from $X \amalg D^2$ to $X \cup_f D^2$. Then let $U_1 = \pi(X \cup D^2 \setminus \{0\})$ and $U_2 = \pi(D^2)$. Then $U_1 \cup U_2 = X \cup_f D^2$, and $U_1 \cap U_2 = (D^2)^\circ \setminus \{0\}$ is path-connected. $\pi_1(U_2) = 1$, so $\pi_1(X \cup_f D^2)$ is generated by $\pi_1(X)$. Note that $f_\star : \pi_1(S^1, 1) \to \pi_1(X, x_0)$, so $f_\star(1)$ lies in the kernel of the inclusion $\pi_1(X, x_0) \to \pi_1(X \cup_f D^2, x_0)$, since $f_\star(1)$ is null-homotopic in $X \cup_f D^2$. So $\pi_1(X \cup_f D^2)$ surjects onto ${}^{\pi_1(X)}\!/_{\langle\langle f_\star(1) \rangle\rangle}$.

This is in fact an isomorphism. Suppose $[\gamma] \in \pi_1(X \cup_f D^2, x_0)$ is mapped to the trivial element of ${}^{\pi_1(X)}\!/_{\langle\langle f_\star(1) \rangle\rangle}$, so $[\gamma]$ can be viewed as an element of $\langle\langle f_\star(1) \rangle\rangle$. Note that all such $[\gamma]$ are of the form $a_1 f_\star(n_1) a_1^{-1} \dots a_k f_\star(n_k) a_k^{-1}$. Since $f_\star(n) = 1$ in $\pi_1(X \cup_f D^2, x_0)$, $[\gamma] = 1$.

5.4. Amalgamated free products

Definition. Let $\iota_1 \colon H \to G_1, \iota_2 \colon H \to G_2$ be group homomorphisms. A group G is an *amalgamated free product of G_1 and G_2 along H* if:

(i) There are homomorphisms $\varphi_1 \colon G_1 \to G, \varphi_2 \colon G_2 \to G$ such that the following diagram commutes.

$$\begin{tikzcd}
 & G_1 \arrow[dr, "\varphi_1"] & \\
H \arrow[ur, "\iota_1"] \arrow[dr, "\iota_2"'] & & G \\
 & G_2 \arrow[ur, "\varphi_2"'] &
\end{tikzcd}$$

(ii) It is universal with this property, so for any other group $\overline{G}$ with a commutative square as above, there is a unique homomorphism $\psi \colon G \to \overline{G}$ such that the following diagram commutes.

$$\begin{tikzcd}
 & G_1 \arrow[dr, "\varphi_1"'] \arrow[drr, bend left, "j_1"] & & \\
H \arrow[ur, "\iota_1"] \arrow[dr, "\iota_2"'] & & G \arrow[r, dashed, "\psi"] & \overline{G} \\
 & G_2 \arrow[ur, "\varphi_2"] \arrow[urr, bend right, "j_2"'] & &
\end{tikzcd}$$

Remark. The amalgamated free product is the colimit of the following diagram.

$$\begin{tikzcd}
 & G_1 \\
H \arrow[ur, "\iota_1"] \arrow[dr, "\iota_2"'] & \\
 & G_2
\end{tikzcd}$$

Hence, it is a categorical pushout.

Proposition. If G, G' are amalgamated products of G_1, G_2, then $G \simeq G'$.

Proof. There are homomorphisms $\alpha \colon G \to G', \beta \colon G' \to G$, and the uniqueness in the definition implies $\alpha \circ \beta = \mathrm{id}_{G'}$ and $\beta \circ \alpha = \mathrm{id}_G$. In other words, the following diagram commutes.

$$\begin{tikzcd}
 & & G_1 \arrow[dl, "\varphi_1"'] \arrow[dr, "\varphi_1'"] & \\
H \arrow[urr, bend left, "\iota_1"] \arrow[drr, bend right, "\iota_2"'] & G \arrow[rr, dashed, shift left, "\alpha"] & & G' \arrow[ll, dashed, shift left, "\beta"] \\
 & & G_2 \arrow[ul, "\varphi_2"] \arrow[ur, "\varphi_2'"'] &
\end{tikzcd}$$

□

Proposition. An amalgamated product of any two groups exists.

The universal property of the presentation is that $\langle S \mid R\rangle \simeq F_S/\langle\!\langle R\rangle\!\rangle$. Suppose $S \subseteq G$ satisfies the relations R in G, so all of the relations map to the identity. Then there is a unique homomorphism $\langle S \mid R\rangle \to G$ mapping $s \in S$ to s, since there is a unique homomorphism $F_S \to G$ mapping $s \in S$ to s, and since S satisfies the relations, this factors through $F_S/\langle\!\langle R\rangle\!\rangle$.

For example, consider a map $\langle a \mid a^4\rangle \to \mathbb{Z}/2\mathbb{Z}$ that maps a to 1. We can check that the relation $1^4 = 0$ in $\mathbb{Z}/2\mathbb{Z}$ holds.

Proof. Consider presentations $G_i = \langle S_i \mid R_i\rangle$ of G_1, G_2, and $H = \langle T \mid W\rangle$. Then define

$$G = G_1 *_H G_2 = \left\langle S_1 \cup S_2 \cup T \mid R_1 \cup R_2 \cup \left\{t_i^{-1}\iota_1(t_i), t_i^{-1}\iota_2(t_i) \mid t_i \in T\right\}\right\rangle$$

Then $\varphi_i \colon G_i \to G$ are given by $s \in S_i$ mapping to s. Given $j_1, j_2 \colon G_1, G_2 \to \overline{G}$, we define $\psi \colon G \to \overline{G}$ mapping $s \in S_1$ to $j_1(s)$, $s \in S_2$ to $j_2(s)$, and $t \in T$ to $j_1 \circ \iota_1(t) = j_2 \circ \iota_2(t)$, and check that the relations hold. □

This is isomorphic to $\left\langle S_1 \cup S_2 \,|\, R_1 \cup R_2 \cup \left\{\iota_1(t_i)\iota_2^{-1}(t_i) \mid t_i \in T\right\}\right\rangle$.

5.5. Seifert–Van Kampen theorem

Theorem (Seifert–Van Kampen). Let $X = U_1 \cup U_2$ where U_i are open sets with $U_1 \cap U_2$ path-connected and containing x_0. Let $G_i = \pi_1(U_i, x_0)$, and $H = \pi_1(U_1 \cap U_2, x_0)$, so

$$\begin{tikzcd}
 & U_1 \arrow[dr, "j_1"] & \\
U_1 \cap U_2 \arrow[ur, "\iota_1"] \arrow[dr, "\iota_2"'] & & X \\
 & U_2 \arrow[ur, "j_2"'] &
\end{tikzcd}
\qquad
\begin{tikzcd}
 & G_1 \arrow[dr, "j_{1\star}"] & \\
H \arrow[ur, "\iota_{1\star}"] \arrow[dr, "\iota_{2\star}"'] & & \pi_1(X) \\
 & G_2 \arrow[ur, "j_{2\star}"'] &
\end{tikzcd}$$

Then $\pi_1(X, x_0) = G_1 *_H G_2$.

Remark. The ‘easy’ part of the proof is that we have a commutative diagram

$$\begin{tikzcd}
& G_1 \arrow[dr, "\psi_1"] \arrow[rrd, bend left, "j_{1\star}"] & & \\
H \arrow[ur, "\iota_{1\star}"] \arrow[dr, "\iota_{2\star}"'] & & G_1 *_H G_2 \arrow[r, dashed, "\psi"] & \pi_1(X) \\
& G_2 \arrow[ur, "\psi_2"] \arrow[rru, bend right, "j_{2\star}"'] & &
\end{tikzcd}$$

so we obtain a map $\psi\colon G_1 *_H G_2 \to \pi_1(X, x_0)$ by universality of the amalgamated free product. Clearly ψ is surjective by the theorem in the previous subsection, and the difficult part of the proof is showing that ψ is injective.

Proof sketch. We show that if $H\colon I \times I \to X$ is a homotopy between γ_0 and γ_1, then $[\gamma_0] = [\gamma_1]$ using the relations in $G_1 *_H G_2$. We can divide $I \times I$ into squares of size $\frac{1}{n}$ such that the image of each square under H lies in either U_1 or U_2 by the Lebesgue covering lemma. Each row represents a path $\gamma_{\frac{i}{n}}$, and by operating row-by-row we will show $\gamma_{\frac{i}{n}}$ is related to $\gamma_{\frac{i+1}{n}}$ in $G_1 *_H G_2$. To move from one row to the next, if there are different labels above and below, the boundary lies in $U_1 \cap U_2$, so we use the relations $\iota_{1\star}(t_1) = \iota_{2\star}(t_1)$. □

Example. Consider $X \cup_f D^2 = U_1 \cup U_2$ where $U_1 = X \cup_f D^2 \setminus \{0\}$ and $U_2 = (D^2)^\circ$, with $x_0 \in U_1 \cap U_2$. Let $p\colon U_1 \to X$ be the inclusion. Since $D^2 \setminus \{0\}$ has a strong deformation retraction to S^1, we know U_1 has a strong deformation retraction to X, so $\pi_1(U_1, x_0) \simeq \pi_1(X, p(x_0))$. Note that $\pi_1(U_2, x_0)$ is the trivial group, since $(D^2)^\circ$ is contractible. Note that $U_1 \cap U_2 = (D^2)^\circ \setminus \{0\}$ is homotopy equivalent to S^1, so $\pi_1(U_1 \cap U_2, x_0) = \mathbb{Z} = \langle \gamma \rangle$.

Then, by the Seifert–Van Kampen theorem, we have $\pi_1(X \cup_f D^2) \simeq \pi_1(X) *_{\mathbb{Z}} 1$. If $\pi_1(X, x_0) = \langle S \mid R \rangle$, we have in particular that

$$\pi_1(X \cup_f D^2) \simeq \left\langle S, t \,\middle|\, R \cup \left\{t, t^{-1} f_\star(t)\right\}\right\rangle = \langle S \mid R \cup f_\star(t) \rangle = {}^{\pi_1(X, x_0)}\!\big/_{\langle\!\langle f_\star(t) \rangle\!\rangle}$$

Example. Consider the torus $T^2 = S^1 \vee S^1 \cup_f D^2$. Let a, b be generators for $\pi_1(S^1 \vee S^1)$. Then the commutator $aba^{-1}b^{-1}$ represents the disk attached. So $\pi_1(T^2) = \left\langle a, b \mid aba^{-1}b^{-1} \right\rangle = \mathbb{Z}^2$.

Example. Let Σ_g be a surface of genus g. Then $\Sigma_g = \bigvee_{i=1}^g (S^1 \vee S^1) \cup_f D^2$, so

$$\pi_1(\Sigma_g) \simeq \left\langle a_1, b_1, \dots, a_g, b_g \,\middle|\, \prod_{i=1}^{g} a_i b_i a_i^{-1} b_i^{-1} \right\rangle$$

Example. A surface of genus two can be realised as a union of U_1, U_2 where $U_1 \cap U_2 \simeq S^1$ and $\pi_1(U_i) = \langle a_i, b_i \rangle$, then $\pi_1(\Sigma_2) = \langle a_1, b_1 \rangle *_{\mathbb{Z}} \langle a_2, b_2 \rangle$.

6. Simplicial complexes

6.1. Simplices

We have shown that $\pi_1(S^1, x_0) \simeq \mathbb{Z}$, and $\pi_1(S^n, x_0) \simeq 1$ for $n > 1$, so $S^1 \nsim S^n$. We would like to show that $S^n \sim S^m$ only holds if $n = m$. One proof of this fact is that any $f : S^n \to S^m$ with $n < m$ is null-homotopic, but the identity on S^m is not. Both of these claims require proof: simplicial complexes will allow us to prove the first, and homology will allow us to prove the second.

Definition. The *n-simplex* is the topological space

$$\Delta^n = \left\{ (x_0, \dots, x_n) \in \mathbb{R}^{n+1} \;\middle|\; x_i \geq 0, \sum_{i=0}^n x_i = 1 \right\}$$

with the subspace topology.

Remark. Δ^1 is homeomorphic to I. Δ^2 is an equilateral triangle, and Δ^3 is a regular tetrahedron. For all n, Δ^n is closed and bounded in $\mathbb{R}^{n+1}$, and hence compact and Hausdorff. The standard basis vectors $e_0, \dots, e_n$ are the vertices of Δ^n.

Definition. If $I \subseteq \{0, \dots, n\}$, the *Ith face of* Δ^n is

$$e_I = \{ x \in \Delta^n \mid x_i = 0 \text{ for } i \notin I \}$$

We define $F(\Delta^n) = \{ e_I \mid I \subseteq \{0, \dots, n\}\}$ to be the set of faces of Δ^n.

If $I = \{i_0, \dots, i_k\}$ with $i_0 < \dots < i_k$, we write $I = i_0 i_1 \dots i_k$.

Remark. Note that $e_{\{i\}} = e_i$, and $\Delta^n = e_{\{0,1,\dots,n\}}$. e_I is a closed subset of Δ^n, and is homeomorphic to $\Delta^{|I|-1}$. $e_I \subseteq e_J$ if and only if $I \subseteq J$. $e_I \cap e_J = e_{I \cap J}$.

Definition. A map $|f| : \Delta^n \to \mathbb{R}^N$ is *affine linear* if it is the restriction of a linear map $\mathbb{R}^{n+1} \to \mathbb{R}^n$. Equivalently, $|f|\big(\sum_{i=0}^n x_i e_i\big) = \sum_{i=0}^n x_i |f|(e_i)$. We say an affine linear map $|f| : \Delta^n \to \Delta^m$ is *simplicial* if it maps vertices in Δ^n to vertices in Δ^m, so there is a map of sets $\hat{f} : \{0, \dots, n\} \to \{0, \dots, m\}$ where $|f|(e_i) = e_{\hat{f}(i)}$.

Remark. Affine linear maps are continuous, and are determined entirely by their action on e_i. In particular, simplicial maps $|f|$ are determined by $\hat{f}$. For $I \subseteq \{0, \dots, n\}$, we have $|f|(e_I) = e_{\hat{f}(I)}$.

Definition. Vectors $v_0, \dots, v_n \in \mathbb{R}^N$ are *affine linearly independent* if whenever $\sum t_i v_i = 0$ and $\sum t_i = 0$, we have $t_i = 0$ for all i. Equivalently,

(i) If $\sum t_i v_i = \sum t_i' v_i$ and $\sum t_i = \sum t_i'$, then for each i, $t_i = t_i'$.

(ii) The vectors $v_1 - v_0, v_2 - v_0, \dots, v_n - v_0$ are linearly independent.

(iii) The unique affine linear map $|f| : \Delta^n \to \mathbb{R}^N$ given by $|f|(e_i) = v_i$ is injective.

If $v_0, \dots, v_n$ are affine linearly independent, we write

$$[v_0, \dots, v_n] = \operatorname{Im} |f| = \left\{ \sum x_i v_i \mid \sum x_i = 1, x_i \geq 0 \right\}$$

and we say $[v_0, \dots, v_n]$ is a *Euclidean simplex.*

Remark. Δ^n is compact and $[v_0, \dots, v_n]$ is Hausdorff, so by the topological inverse function theorem, $|f| : \Delta^n \to [v_0, \dots, v_n]$ is a homeomorphism if the v_i are affine linearly independent.

Lemma. If $X \subseteq \mathbb{R}^N$, let $Z(X)$ be the set of $x \in X$ such that if $x = \sum t_i x_i$ for $t_i > 0, \sum t_i = 1$ and all $x_i \in X$, then $x_i = x$ for some i. Then $Z([v_0, \dots, v_n]) = \{v_0, \dots, v_n\}$.

Proof. We show that $v_k \in Z([v_0, \dots, v_n])$; the converse is clear from the definition of the simplex. Suppose $v_k = \sum t_i x_i$ for $t_i > 0$ and $\sum t_i = 1$. Then $x_i = \sum_{j=0}^n s_{ij} v_j$, since $x_i \in [v_0, \dots, v_n]$. So $v_k = \sum_j \left(\sum_i t_i s_{ij} \right) v_j$. Since the v_i are affine linearly independent, and $\sum_j \left(\sum_i t_i s_{ij} \right) = 1$, we must have $\sum t_i s_{ij} = 0$ for $j \neq k$. But $t_i > 0$ and $s_{ij} \geq 0$, so the only case is when all s_{ij} are exactly zero for $j \neq k$, so $x_j = v_k$. □

Corollary. If $[v_0, \dots, v_n] = [v_0', \dots, v_n']$ as subsets of $\mathbb{R}^N$, then $\{v_0, \dots, v_n\} = \{v_0', \dots, v_n'\}$ as sets.

Therefore, a simplex determines its set of vertices.

Proof. $\{v_0, \dots, v_n\} = Z([v_0, \dots, v_n]) = Z([v_0', \dots, v_n']) = \{v_0', \dots, v_n'\}$. □

Definition. $\mathcal{S}(\mathbb{R}^n)$ is the set of Euclidean simplices $\sigma \subseteq \mathbb{R}^n$. Hence, $\mathcal{S}(\mathbb{R}^n)$ is in bijection with the set $\left\{ \{v_0, \dots, v_k\} \mid v_i \in \mathbb{R}^N, k \geq -1, v_i \text{ affine linearly independent} \right\}$.

6.2. Abstract simplicial complexes

Definition. An *abstract simplicial complex* in Δ^n is a subset K of the faces $F(\Delta^n)$ such that $e_I \in K$ whenever e_J is in K and $I \subseteq J$.

Remark. Abstract simplicial complexes are downward-closed sets of faces. They have no intrinsic topology. The set of faces $F(\Delta^n)$ of the n-dimensional simplex Δ^n is an abstract simplicial complex.

Definition. If K is an abstract simplicial complex, its *polyhedron* is $|K| = \bigcup_{e_I \in K} e_I \subseteq \Delta^n$.

Remark. Polyhedra are compact and Hausdorff.

Definition. We define $K_r = \{e_I \in K \mid |I| \leq r + 1\}$ to be the set of faces of dimension at most r. This is called the *r-skeleton* of K.

The r-skeleton is an abstract simplicial complex. Note that

$$\{e_\varnothing\} = K_{-1} \subset K_0 \subset \cdots \subset K_n = K$$

We write $\dim K = \max\{\dim e_I \mid e_I \in K\}$.

Definition. The *vertex set* $V(K)$ is the polyhedron $|K_0|$.

Example. $\mathbf{\Delta}^n = F(\Delta^n) = \{e_I \mid I \subseteq \{0, \dots, n\}\}$ is a simplicial complex. Its polyhedron is Δ^n, which is homeomorphic to D^n by radial projection.

Example. $\mathbb{S}^{n-1} = \mathbf{\Delta}^n_{n-1} = \{e_i \mid I \subsetneq \{0, \dots, n\}\}$ is a simplicial complex. This has polyhedron $\partial\Delta^n$ by definition of the boundary. This is homeomorphic to S^{n-1} by radial projection.

Definition. Let K, L be abstract simplicial complexes in Δ^n and Δ^m respectively. A *simplicial map* $f : K \to L$ is a map such that there is a simplicial map $|f| : \Delta^n \to \Delta^m$ with $f(e_I) = |f|(e_I)$. Equivalently, there is a map $\hat{f} : \{0, \dots, n\} \to \{0, \dots, m\}$ such that $f(e_I) = e_{\hat{f}(I)}$ and $e_I \in K$ implies $e_{\hat{f}(I)} \in L$.

Remark. The identity map is simplicial. The composition of two simplicial maps is simplicial.

Definition. We say a simplicial map $f : K \to L$ is a *simplicial isomorphism* if f is a bijection, or equivalently, $|f|$ is a bijection or $|f|$ is a homeomorphism, treating $|f|$ as a map $|K| \to |L|$.

6.3. Euclidean simplicial complexes

Recall that $\mathcal{S}(\mathbb{R}^n)$ is the set of Euclidean simplices $[v_0, \dots, v_n]$ where the v_i are affine linearly independent.

Definition. $K \subseteq \mathcal{S}(\mathbb{R}^n)$ is a *Euclidean simplicial complex* if

(i) K is finite;

(ii) if $\sigma \in K$ and $\tau \in F(\sigma)$, then $\tau \in K$;

(iii) if $\sigma_1, \sigma_2 \in K$, then $\sigma_1 \cap \sigma_2 \in F(\sigma_1) \cap F(\sigma_2)$, so in particular, $\sigma_1 \cap \sigma_2 \in K$.

If so, we write $|K| = \bigcup_{\sigma \in K} \sigma \subseteq \mathbb{R}^n$ with the subspace topology. We write

$$K_r = \{\sigma \in K \mid \dim \sigma \leq r\}$$

for its r-skeleton, which is a Euclidean simplicial complex.

Proposition. Let $|\varphi| : \Delta^n \to \mathbb{R}^n$ be affine linear, and K' be an abstract simplicial complex in Δ^n, such that $|\varphi||_{|K'|}$ is injective. Then $\varphi(K') = \{|\varphi|(e_I) \mid e_I \in K'\}$ is a Euclidean simplicial complex.

Proof. Property (i) is clear since $F(\Delta^n)$ is finite. For property (ii), note that if $\sigma \in \varphi(K')$, there is $e_I \in K'$ such that $\sigma = |\varphi|(e_I)$. If $\tau \in F(\sigma)$, we have $\tau = |\varphi|(e_J)$ for $e_J \subseteq e_I$. Then $e_J \in K'$ since K' is an abstract simplicial complex. So $\tau = |\varphi|(e_J) = \varphi(K')$.

Suppose $\sigma_1 = |\varphi|(e_{I_1})$ and $\sigma_2 = |\varphi|(e_{I_2})$ where $e_{I_1}, e_{I_2} \in K'$. Then $\sigma_1 \cap \sigma_2 = |\varphi|(e_{I_1}) \cap |\varphi|(e_{I_2}) = |\varphi|(e_{I_1} \cap e_{I_2})$ by injectivity. This is equal to $|\varphi|(e_{I_1 \cap I_2}) \in F(\sigma_1) \cap F(\sigma_2)$. □

Definition. We say that the Euclidean simplicial complex $\varphi(K')$ is a *realisation* of an abstract simplicial complex K' in Δ^n, if $|\varphi| : \Delta^n \to \mathbb{R}^n$ is affine linear and injective on $|K'|$.

Remark. If $\varphi(K')$ is a realisation of K', $|\varphi||_{|K'|}$ is injective, so $|\varphi| : |K'| \to |\varphi(K)|$ is a homeomorphism.

Proposition. Let $K \subseteq \mathbb{R}^N$ be a Euclidean simplicial complex. Then $K = \varphi(K')$ for some abstract simplicial complex K', and $|\varphi| : |K'| \to |K|$. Any two K' are related by a simplicial isomorphism.

Informally, every Euclidean simplicial complex is the realisation of some abstract simplicial complex.

Proof. Let $V(K) = |K_0| = \{v_0, \dots, v_n\} \subset \mathbb{R}^N$ be the vertex set of the Euclidean simplicial complex. Define $K' = \left\{e_{\{i_0,\dots,i_k\}} \mid [v_{i_0}, \dots, v_{i_k}] \in K\right\}$. Let $|\varphi| : \Delta^n \to \mathbb{R}^N$ be given by $|\varphi|(e_i) = v_i$.

We show that $|\varphi||_{|K'|}$ is injective. If $\sigma = [v_{i_0}, \dots, v_{i_k}] \in K$, we have that $v_{i_0}, \dots, v_{i_k}$ are affine linearly independent since K is a Euclidean simplicial complex. Then $|\varphi|_{e_I}$ is injective.

Suppose $|\varphi|(p) = |\varphi|(q) = x \in \mathbb{R}^N$, where $p \in e_I \in K'$ and $q \in e_J \in K'$. Then $x \in |\varphi|(e_I) \cap |\varphi|(e_J)$, which is the intersection of simplices in K, so $x \in |\varphi|(e_{I'})$ for $I' \subseteq I \cap J$. Since $|\varphi||_{e_I}$ and $|\varphi||_{e_J}$ are injective, we must have $p, q \in e_{I'}$. But $|\varphi||_{e_{I'}}$ is also injective, so $p = q$. □

Definition. A *simplicial map of Euclidean simplicial complexes* is a map $f : K_1 \to K_2$ if there are realisations $\varphi_i : K_i' \to K_i$ and a simplicial map of abstract simplicial complexes $f' : K_1' \to K_2'$ so that the following diagram commutes.

$$\begin{tikzcd}
K_1' \arrow[r, "f'"] \arrow[d, "\varphi_1"'] & K_2' \arrow[d, "\varphi_2"] \\
K_1 \arrow[r, "f"'] & K_2
\end{tikzcd}$$

Remark. The composition of simplicial maps of Euclidean simplicial complexes is also a simplicial map.

6.4. Boundaries and cones

Definition. Let σ be an n-dimensional Euclidean simplex. Let $F(\sigma)$ be the set of faces of σ, a Euclidean simplicial complex with $|F(\sigma)| = \sigma$. Let $\not\partial\sigma = F(\sigma)_{n-1} = F(\sigma) \setminus \{\sigma\}$, a Euclidean simplicial complex. Let $\partial\sigma = |\not\partial\sigma| \subset \mathbb{R}^N$ be the boundary of σ. It is homeomorphic to S^{n-1}. Let $\sigma^\circ = \sigma \setminus \partial\sigma$ be the interior of σ.

Definition. Let $X \subseteq \mathbb{R}^N$ and $p \in \mathbb{R}^N$. We say p is *independent* of X if for each $x \in X$, the ray px from p to x has $px \cap X = \{x\}$.

Definition. If p is independent of X, the *cone* is defined by

$$C_p(X) = \{tp + (1-t)x \mid t \in [0,1], x \in X\}$$

Example. Let $X = [v_0, \dots, v_n]$ be an n-simplex. Then p is independent of X if and only if $\{v_0, \dots, v_n, p\}$ is an affine linearly independent set. If so, $C_p(X) = [v_0, \dots, v_n, p]$.

Definition. Let K be a Euclidean simplicial complex in $\mathbb{R}^N$ and p be independent of $|K|$. Then we define the *cone*

$$C_p(K) = K \cup \big\{[v_0, \dots, v_j, p] \mid [v_0, \dots, v_j] \in K\big\}$$

Lemma. If p is independent of $|K|$, then $C_p(K)$ is a Euclidean simplicial complex and $\left|C_p(K)\right| = C_p(|K|)$.

6.5. Barycentric subdivision

Definition. If $\sigma = [v_0, \dots, v_n]$ is an n-simplex in $\mathbb{R}^N$, we define its *barycentre*

$$b_\sigma = \frac{1}{n+1} \sum_{i=0}^{n} v_i$$

Lemma. b_σ is independent of $\partial\sigma$, and $C_{b_\sigma}(\partial\sigma) = \sigma$.

We will define maps β from $\mathcal{S}(\mathbb{R}^N)$ to the set of Euclidean simplicial complexes in $\mathbb{R}^N$, and B from the set of Euclidean simplicial complexes in $\mathbb{R}^N$ to Euclidean simplicial complexes in $\mathbb{R}^N$, satisfying $|\beta(\sigma)| = \sigma$ and $|B(K)| = |K|$. The maps β and B are called *barycentric subdivision*. In order to do this, we will inductively define β and B on simplices and Euclidean simplicial complexes of dimension at most n, and prove the following theorems.

Theorem (first inductive hypothesis)**.** Let $\sigma \in \mathcal{S}(\mathbb{R}^N)$ be an n-simplex. Then $\beta(\sigma)$ is a Euclidean simplicial complex of dimension n, and $|\beta(\sigma)| = \sigma$. If τ is a face of σ and $\sigma_1 \in \beta(\sigma)$ then $\sigma_1 \cap \tau \in \beta(\tau)$.

Theorem (second inductive hypothesis)**.** Let K be an n-dimensional Euclidean simplicial complex. Then $B(K)$ is an n-dimensional Euclidean simplicial complex with polyhedron $|B(K)| = |K|$.

For the base case, let $n = -1$. The only -1-dimensional simplex is $\varnothing$. We define $\beta(\varnothing) = \{\varnothing\}$. The only -1-dimensional simplicial complex is $\{\varnothing\}$, and we define $B(\{\varnothing\}) = \{\varnothing\}$. Both inductive hypotheses hold for this case.

In general, suppose β and B are defined on $n-1$-dimensional simplices and simplicial complexes and that both inductive hypotheses hold. We now define $\beta(\sigma) = C_{b_\sigma}(B(\not\partial\sigma))$ and $B(K) = \bigcup_{\sigma \in K} \beta(\sigma)$.

Example. Let σ be a zero-dimensional simplex. Then $\beta_\sigma(\sigma) = \sigma$.

Example. Let σ be the one-dimensional simplex. $\not\partial\sigma$ is two points p_1, p_2 and the empty set. Then $B(\not\partial\sigma) = \{\varnothing, p_1, p_2\}$. Therefore, $C_p(B(\not\partial\sigma)) = \{\varnothing, p, p_1, p_2, pp_1, pp_2\}$.

Example. Let σ be a two-dimensional simplex with vertices p_1, p_2, p_3. Then $C_p(B(\not\partial\sigma))$ has six 2-simplices, twelve 1-simplices, seven 0-simplices and one empty simplex.

Proof of first inductive hypothesis. $\not\partial\sigma$ is a Euclidean simplicial complex of dimension $n - 1$, hence $B(\not\partial\sigma)$ is a Euclidean simplicial complex by the second inductive hypothesis, and $|B(\not\partial\sigma)| = |\not\partial\sigma| = \partial\sigma$. By the lemmas above, b_σ is independent of $\partial\sigma = |B(\not\partial\sigma)|$, so $C_{b_\sigma}(B(\not\partial\sigma))$ is a Euclidean simplicial complex with polyhedron $\left|C_{b_\sigma}(B(\not\partial\sigma))\right| = C_{b_\sigma}(\partial\sigma) = \sigma$. The next part follows from the lemma: if $\sigma \in C_p(K)$, then $\sigma \cap |K| \in K$. □

Proof of second inductive hypothesis. We check the properties required for a Euclidean simplicial complex for $B(K) = \bigcup_{\sigma\in K} \beta(\sigma)$. $\beta(\sigma)$ is finite for each σ and K is finite, so $B(K)$ is finite. If $\sigma \in B(K)$ then $\sigma \in \beta(\sigma')$ for some $\sigma' \in K$, so if $\tau \in F(\sigma)$, then $\tau \in \beta(\sigma')$ since $\beta(\sigma')$ is a Euclidean simplicial complex, so $\tau \in B(K)$, so the second property holds. Suppose $\sigma_1, \sigma_2 \in B(K)$ where $\sigma_i \in \beta(\sigma_i')$ and $\sigma_i' \in K$. Then $\sigma_1 \cap \sigma_2 \subseteq \sigma_1' \cap \sigma_2' = \tau$ since $|\beta(\sigma_i')| = \sigma_i'$, where $\tau \in K$ since K is a Euclidean simplicial complex. Then $\sigma_1 \cap \tau, \sigma_2 \cap \tau \in \beta(\tau)$ by the second part of the first inductive hypothesis. In particular, $\beta(\tau)$ is a Euclidean simplicial complex, so $\sigma_1 \cap \sigma_2 = \underbrace{(\sigma_1 \cap \tau)}_{\in\beta(\tau)} \cap \underbrace{(\sigma_2 \cap \tau)}_{\in\beta(\tau)} \in \beta(\tau) \subseteq B(K)$, so the third property holds. So K is a Euclidean simplicial complex. Now, by the first inductive hypothesis, $|B(K)| = \bigcup_{\sigma\in K} \beta(\sigma) = \bigcup_{\sigma\in K} \sigma = |K|$. □

Lemma. Let $\sigma \in \mathcal{S}(\mathbb{R}^N)$ and $x, v \in \sigma$. Then $\|v - x\| \leq \max_{v_i\in V(\sigma)} \|v - v_i\|$.

Proof. We can write $x = \sum x_i v_i$, where $\sum x_i = 1, x_i \geq 0$, and $v_i \in V(\sigma)$. But also, $v = \sum x_i v$. Hence,

$$\|v - x\| = \left\|\sum x_i(v - v_i)\right\| \leq \sum x_i \|v - v_i\| \leq \sum x_i \max \|v - v_i\| = \max \|v - v_i\|$$

□

Applying this twice, $\|x - v\| \leq \max_{v_i\in V(\sigma)} \|v - v_i\| \leq \max_{v_i,v_j\in V(\sigma)} \left\|v_i - v_j\right\|$.

Definition. The *mesh* of a simplex $\sigma \in \mathcal{S}(\mathbb{R}^N)$ is

$$\mu(\sigma) = \max_{v_i,v_j\in V(\sigma)} \left\|v_i - v_j\right\| = \max_{x,v\in\sigma} \|v - x\|$$

If K is a Euclidean simplicial complex, its mesh is $\mu(K) = \max_{\sigma\in K} \mu(\sigma)$.

Lemma. Let b_σ be the barycentre of σ, so $b_\sigma = \frac{1}{n+1}\sum_{i=0}^n v_i$ for $\sigma = [v_0, \dots, v_n]$. Then $\max_{v\in\sigma} \|b_\sigma - v\| \leq \frac{n}{n+1}\mu(\sigma)$.

Proof. $\|b_\sigma - v\| \leq \max_{v_i \in V(\sigma)} \|b_\sigma - v_i\|$. We have

$$\|b_\sigma - v_i\| = \frac{1}{n+1}\left\|\sum_{j\neq i} v_j - nv_i\right\| \leq \frac{1}{n+1}\sum_{j\neq i}\left\|v_j - v_i\right\| \leq \frac{1}{n+1}\cdot n\mu(\sigma)$$

□

Corollary. Let σ be a Euclidean simplex of dimension n. Then $\mu(\beta(\sigma)) \leq \frac{n}{n+1}\mu(\sigma)$. Let K be a Euclidean simplicial complex of dimension n. Then $\mu(B(K)) \leq \frac{n}{n+1}\mu(K)$.

Proof. Let $\tau \in \beta(\sigma)$. Suppose $\tau \in B(\partial\sigma)$. Then, $\mu(\tau) \leq \frac{n-1}{n}\mu(B(\partial\sigma)) \leq \frac{n}{n+1}\mu(\sigma)$ by induction. Otherwise, $\tau = [v_0, \dots, v_k, b_\sigma]$, where $[v_0, \dots, v_k] \in B(\partial\sigma)$. Then $\left\|v_i - v_j\right\| \leq \frac{n-1}{n}\mu(\sigma)$ by induction, and $\|v_i - b_\sigma\| \leq \frac{n}{n+1}\mu(\sigma)$ by the lemma. □

6.6. Simplicial approximation

Lemma. (i) Let $x \in \Delta^n$. Then there exists a unique $I \subseteq \{0, \dots, n\}$ such that $x \in e_I^\circ$.

(ii) If $x \in e_I^\circ$, then $x \in e_J$ if and only if $I \subseteq J$, or equivalently, $e_I \subseteq e_J$.

(iii) Let K be an abstract simplicial complex in Δ^n, and let $x \in e_I^\circ$. Suppose that $x \in |K|$. Then $e_I \in K$.

Proof. *Part (i).* Let $I = \{i \in \{0, \dots, n\} \mid x_i \neq 0\}$. *Part (ii).* Follows from part (i).

Part (iii). $x \in |K|$ implies $x \in e_J$ for some $e_J \in K$. By part (ii), we have $e_I \subseteq e_J$. Since K is an abstract simplicial complex and $e_J \in K$, we have $e_I \in K$. □

Corollary. Let K be a Euclidean simplicial complex, and $x \in |K|$. Then there exists a unique $\sigma \in K$ with $x \in \sigma^\circ$.

Proof. Let $\varphi : K' \to K$ be a realisation of K, so K' is an abstract simplicial complex and φ is a bijection inducing a homeomorphism on the polyhedra. Let $x' = |\varphi^{-1}|(x) \in |K|$. Then x' lies in the interior of a unique e_I by part (i) of the lemma above. Note that $e_I \in K'$ by part (iii), so $\varphi(e_I)$ is the unique $\sigma \in K$ with $x \in \sigma^\circ$. □

Definition. Let K be a Euclidean simplicial complex, and let $v \in V(K)$. Then the *star* $\mathrm{St}_K(v)$ is $\bigcup_{\{\sigma\in K \mid v\in\sigma\}} \sigma^\circ$.

Lemma. (i) Let $x \in |K|$ and $x \in \sigma^\circ$. Then $x \in \mathrm{St}_K(v)$ if and only if $v \in V(\sigma)$.

(ii) $\mathrm{St}_K(v) = |K| \setminus \bigcup_{\{\sigma\in K \mid v\notin V(\sigma)\}} \sigma^\circ = |K| \setminus \bigcup_{\{\sigma\in K \mid v\notin V(\sigma)\}} \sigma$.

(iii) $\{\mathrm{St}_K(v) \mid v \in V(K)\}$ is an open cover of $|K|$.

Proof. Part (i). Follows from the fact that if $x \in |K|$, x lies in a unique interior of σ for $\sigma \in K$.

Part (ii). The first equality follows from part (i). The second follows from the fact that if $\tau \in F(\sigma)$ and $v \notin V(\sigma)$, then $v \notin V(\tau)$.

Part (iii). Part (ii) exhibits $\mathrm{St}_K(v)$ as the complement of a finite union of closed sets in $|K|$, so it is open. If $x \in |K|$, then $x \in \sigma^\circ$ for some σ, and if $v \in V(\sigma)$, then $x \in \mathrm{St}_K(v)$, so it is a cover. □

Definition. Let K, L be Euclidean simplicial complexes. Let $f : |K| \to |L|$ be a continuous map, and let $\hat{g} : V(K) \to V(L)$. We say that $\hat{g}$ is a *simplicial approximation* of f if $f(\mathrm{St}_K(v)) \subseteq \mathrm{St}_L(\hat{g}(v))$ for all $v \in V(K)$.

Theorem. Let $\varphi : K' \to K$ be a realisation of a Euclidean simplicial complex K, and let L be a Euclidean simplicial complex in $\mathbb{R}^M$. We define $g' : |K'| \to \mathbb{R}^M$ to be the affine linear map with $|g'|(v) = \hat{g}(\varphi(v))$ if $v \in V(K')$. Let $|g| = |g'| \circ |\varphi|^{-1}$. Then $|g|$ defines a simplicial map $g : K \to L$, and $|g| \sim f$.

Proof. Let $\sigma \in K$. We must show that $|g|(\sigma) \in L$. Let $x \in \sigma^\circ$ be an arbitrary point in the interior. Then $f(x) \in |L|$, so $f(x) \in \tau^\circ$ with $\tau \in L$. Then $x \in \bigcap_{v \in V(\sigma)} \mathrm{St}_K(v)$, so $f(x) \in \bigcap_{v \in V(\sigma)} f(\mathrm{St}_K(v)) \subseteq \bigcap_{v \in V(\sigma)} \mathrm{St}_L(g(v))$ since g is a simplicial approximation of f. Now, if $v \in V(\sigma)$, $f(x) \in \tau^\circ$ and $f(x) \in \mathrm{St}_L(g(v))$, so $g(v) \in \tau$ by part (i) of the lemma above. Hence, every vertex of $|g|(\sigma)$ is a vertex of τ, so $|g|(\sigma)$ is a face of $\tau \in L$, so $|g|(\sigma) \in L$ as required. So $g : K \to L$ is simplicial.

For the second part, we define $H : |K| \times I \to \mathbb{R}^M$ by $H(x,t) = t|g|(x) + (1-t)f(x)$. This is clearly a homotopy in $\mathbb{R}^M$, but we need to show it is a homotopy in $|L|$. Suppose $x \in \sigma^\circ$ and $f(x) \in \tau^\circ$ as before. Then $x = \sum_{v_i \in V(\sigma)} x_i v_i$, so $|g|(x) = \sum_{v_i \in V(\sigma)} x_i |g|(v_i) \in \tau$ since $|g|(v_i) \in \tau$. Since τ is convex, and $|g|(x), f(x) \in \tau$, we must have $H(x,t) \in \tau$ for $t \in [0,1]$. So $H : |K| \times I \to |L|$, which is the desired homotopy. □

Theorem (simplicial approximation theorem)**.** Let K, L be Euclidean simplicial complexes. Let $f : |K| \to |L|$ be a continuous map. Then there exists $r > 0$ and a simplicial map $g : B^r(K) \to L$ such that $|g| \sim f$.

Note that $|B^r(K)| = |K|$, so $|g| : |B^r(K)| \to |L|$ can be thought of as a map $|K| \to |L|$.

Proof. We have the open cover $\{\mathrm{St}_L(v) \mid v \in V(L)\}$ of $|L|$. $f : |K| \to |L|$ is continuous, so $\left\{f^{-1}(\mathrm{St}_L(v)) \mid v \in V(L)\right\}$ is an open cover of $|K|$. Now, $|K|$ is a compact metric space, so we can apply the Lebesgue covering lemma to find $\delta > 0$ and a function $|K| \to V(L)$ mapping x to some vertex v_x such that $B_\delta(x) \subseteq f^{-1}(\mathrm{St}_L(v_x))$. Let r be a natural number such that $\mu(B^r(K)) < \delta$, and let $K' = B^r(K)$. If $\sigma \in K'$ and $x \in V(\sigma)$, then $\sigma \subseteq B_\delta(x)$, since $\mu(K') < \delta$. If $x \in V(K')$, then

$$\mathrm{St}_{K'}(x) = \bigcup_{\{\sigma \mid x \in V(\sigma)\}} \sigma^\circ \subseteq \bigcup_{\{\sigma \mid x \in V(\sigma)\}} \sigma \subseteq B_\delta(x)$$

Hence, $f(\mathrm{St}_{K'}(x)) \subseteq f(B_\delta(x)) \subseteq \mathrm{St}_L(v_x)$, so the function $\hat{g} : V(K') \to V(L)$ given by $\hat{g}(x) = v_x$ is a simplicial approximation of f. So by the previous theorem, $\hat{g}$ determines a simplicial map $g : K' \to L$ with $|g| \sim f$. □

Corollary. Let K, L be Euclidean simplicial complexes, where $\dim K < \dim L$. Let $f : |K| \to |L|$ be continuous. Then $f \sim |g|$ where $|g|$ is not surjective.

Proof. Let $g : B^r(K) \to L$ be a simplicial map such that $f \sim |g|$. Let $k = \dim B^r(K) = \dim K$. Then $|g| : |K| \to |L_k| \subsetneq |L|$ since $\dim L > k$. So $|g|$ is not surjective. □

Remark. It is a general fact that simplicial functions map an i-skeleton into an i-skeleton for each i.

Theorem. If $k < n$, any continuous map $S^k \to S^n$ is null-homotopic.

Proof. $S^k \simeq |\mathbb{S}^k|$ and $S^n \simeq |\mathbb{S}^n|$. By the above corollary, $f \sim |g|$ where $|g| : S^k \to S^n$ is not surjective. Let $|g| : S^k \to S^n \setminus \{p\}$.

$$\begin{tikzcd}
S^k \arrow[r, "g'"] \arrow[dr, "|g|"'] & S^n \setminus \{p\} \arrow[d, "\iota"] \\
 & S^n
\end{tikzcd}$$

But $S^n \setminus \{p\} \simeq \mathbb{R}^n$ is contractible. So g' is null-homotopic, so $|g| \sim \iota \circ g'$ is null-homotopic. □

7. Simplicial homology

7.1. Chain complexes

Definition. A *(finitely generated) chain complex* $(C_\bullet, d)$ is

(i) a collection of free (finitely generated) abelian groups C_i for $i \in \mathbb{Z}$ (and if finitely generated, $C_i = 0$ for all but finitely many i);

(ii) a collection of homomorphisms $d_i : C_i \to C_{i-1}$;

(iii) $d_{i-1} \circ d_i = 0$ for all i.

$$\begin{tikzcd} \cdots & C_{-2} \arrow[l, "d_{-2}"'] & C_{-1} \arrow[l, "d_{-1}"'] & C_0 \arrow[l, "d_0"'] & C_1 \arrow[l, "d_1"'] & C_2 \arrow[l, "d_2"'] & \cdots \arrow[l, "d_3"'] \end{tikzcd}$$

Usually, we write $C_\bullet = \bigoplus_i C_i$, and $d = \bigoplus_i d_i : C_\bullet \to C_\bullet$. We can check that $d_{i-1} \circ d_i = 0$ for all i is equivalent to the statement that $d \circ d = d^2 = 0$.

Remark. Free finitely generated abelian groups are isomorphic to $\mathbb{Z}^n$ for some n. A chain complex defined over $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{F}_p$ is similar, except that C_i is a vector space over the $\mathbb{Q}, \mathbb{R}, \mathbb{F}_p$ and the d_i are linear maps. Every chain complex determines another chain complex over $\mathbb{Q}, \mathbb{R}, \mathbb{F}_p$ by replacing $\mathbb{Z}^{n_i}$ with $\mathbb{Q}^{n_i}$, for example, and the d_i are given by the same matrices.

Remark. There is a unique group homomorphism to and from the trivial abelian group 0. Arrows to and from this group can therefore be unlabelled.

Example (reduced chain complex of the simplex)**.** Consider the reduced chain complex of Δ^n. We define $\widetilde{C}_k(\Delta^n) = \langle e_I \mid |I| = k+1, I \subseteq \{0, \dots, n\}\rangle$, the free abelian group on a basis given by the e_I. We also define $d(e_I) = \sum_{j=0}^{|I|} (-1)^j e_{I_{\hat\jmath}}$ where if $I = i_0 i_1 \dots i_k$ and $i_0 < \dots < i_k$, we define $I_{\hat\jmath} = I \setminus \{i_j\}$. For example, consider $\widetilde{C}_\bullet(\Delta^2)$.

$$\widetilde{C}_2(\Delta^2) = \langle e_{012} \rangle; \quad \widetilde{C}_1(\Delta^2) = \langle e_{01}, e_{02}, e_{12} \rangle; \quad \widetilde{C}_0(\Delta^2) = \{e_0, e_1, e_2\}; \quad \widetilde{C}_{-1}(\Delta^2) = \{e_\varnothing\}$$

and, for example,

$$d(e_{012}) = (-1)^0 e_{12} + (-1)^1 e_{02} + (-1)^2 e_{01} = e_{12} - e_{02} + e_{01}$$

$$d(e_{01}) = e_1 - e_0; \quad d(e_{02}) = e_2 - e_0; \quad d(e_{12}) = e_2 - e_1; \quad d(e_0) = d(e_1) = d(e_2) = e_\varnothing$$

Note that $\widetilde{C}_i(\Delta^2) = 0$ if $i < -1$ or $i > 2$. We have $d^2(e_{012}) = d(e_{12} - e_{02} + e_{01}) = e_2 - e_1 - e_2 + e_0 + e_1 - e_0 = 0$, as required.

$$\begin{tikzcd} 0 & \widetilde{C}_{-1} \arrow[l] & \widetilde{C}_0 \arrow[l, "d_0"'] & \widetilde{C}_1 \arrow[l, "d_1"'] & \widetilde{C}_2 \arrow[l, "d_2"'] & 0 \arrow[l] \end{tikzcd}$$

Proposition. For $\widetilde{C}_\bullet(\Delta^n)$, $d^2 = 0$.

Proof. The e_I are a basis for $\widetilde{C}_\bullet(\Delta^n)$, so it suffices to check that $d^2(e_I) = 0$ for each I. For some $c_{jj'}$, we have $d^2(e_I) = \sum_{j<j'} c_{jj'} e_{I_{\hat\jmath,\hat\jmath'}}$ where $I_{\hat\jmath,\hat\jmath'} = I \setminus \{i_j, i_{j'}\}$. We can compute that $c_{jj'}$ has a contribution of $(-1)^j(-1)^{j'-1}$ by first considering j then j', since $i_{j'}$ is the $(j'-1)$th

element of I_j. Note also that by computing the term in the sum with j, j' in the other order, we have a contribution of $(-1)^{j'}(-1)^i$. Hence $c_{jj'} = (-1)^j(-1)^{j'-1} + (-1)^{j'}(-1)^i = 0$. □

Example (chain complex of the simplex)**.** The chain complex of Δ^n is defined by $C_i(\Delta^n) = \widetilde{C}_i(\Delta^n)$ if $i \geq 0$, but $C_{-1}(\Delta^n) = 0$. This removes the empty face $e_\varnothing$. The d_i are unchanged.

$$\begin{tikzcd} 0 & C_0 \arrow[l] & C_1 \arrow[l, "d_1"'] & C_2 \arrow[l, "d_2"'] & 0 \arrow[l] \end{tikzcd}$$

Definition. Let K be an abstract simplicial complex in Δ^n. Let

$$\widetilde{C}_k(K) = \langle e_I \mid |I| = k+1, e_I \in K \rangle \leq \widetilde{C}_k(\Delta^n)$$

Since $e_I \in K$ implies $e_{I_j} \in K$, $d_k : \widetilde{C}_k(K) \to \widetilde{C}_{k-1}(K)$. So $(\widetilde{C}_\bullet(K), d)$ is a chain complex.

Definition. Let $(C_\bullet, d)$ be a chain complex, and let $x \in C_k$. We say that x is a *cycle* or *closed* if $dx = 0$, so $x \in \ker d_k$. We say that x is a *boundary* or *exact* if $x = dy$ for some y, so $x \in \operatorname{Im} d_{k+1}$.

Remark. The statement $d^2 = 0$ is equivalent to the statement $\operatorname{Im} d_{k+1} \subseteq \ker d_k$ for each k, so boundaries are always cycles.

7.2. Homology groups

Definition. Let $(C_\bullet, d)$ be a chain complex. Its *kth homology group* is

$$H_k(C) = {}^{\ker d_k}\!/\!_{\operatorname{Im} d_{k+1}}$$

Remark. Homology groups are abelian.

Example. Consider $\widetilde{C}_\bullet(\Delta^2)$. Recall $\widetilde{C}_2 = \langle e_{012} \rangle$ and $d(e_{012}) = e_{12} - e_{02} + e_{01}$. Hence $\ker d_2 = 0$ and $\operatorname{Im} d_3 = 0$, so $H_2(\widetilde{C}_\bullet(\Delta^2)) = 0$.

We have $\widetilde{C}_1 = \langle e_{12}, e_{02}, e_{01} \rangle$, and $d(ae_{01} + be_{12} + ce_{02}) = a(e_1 - e_0) + b(e_2 - e_1) + c(e_2 - e_0) = -(a+c)e_0 + (a-b)e_1 + (b+c)e_2$. Hence $ae_{01} + be_{12} + ce_{02} \in \ker d$ if and only if $a = b = -c$. So $x \in \langle e_{12} - e_{02} + e_{01} \rangle = \operatorname{Im} d_2$, giving $H_1(\widetilde{C}_\bullet(\Delta^2)) = 0$.

We have $\widetilde{C}_0 = \langle e_0, e_1, e_2 \rangle$ and $d(e_i) = e_\varnothing$, so $\ker d_0 = \{ae_0 + be_1 + ce_2 \mid a + b + c = 0\}$. Conversely, $\operatorname{Im} d_1 = \operatorname{span}\{e_1 - e_0, e_2 - e_0, e_2 - e_1\} = \ker d_0$. So in fact $H_0(\widetilde{C}_\bullet(\Delta^2)) = 0$.

Now $\widetilde{C}_{-1} = \langle e_\varnothing \rangle = \ker d_{-1} = \langle e_\varnothing \rangle = \operatorname{Im} d_0$ so $H_{-1}(\widetilde{C}_\bullet(\Delta^2)) = 0$. So all of the homology groups of $\widetilde{C}_\bullet(\Delta^2)$ are trivial. Note that

$$H_i(C_\bullet(\Delta^2)) = \begin{cases} H_i(\widetilde{C}_\bullet(\Delta^2)) & i > 0 \\ {}^{\langle e_0, e_1, e_2 \rangle}\!/\!_{\operatorname{span}\{e_1 - e_0, e_2 - e_0, e_2 - e_1\}} \simeq \mathbb{Z} & i = 0 \end{cases}$$

Definition. Let K be an abstract simplicial complex in Δ^n. Then we define the ith reduced homology group of K to be $\widetilde{H}_i(K) = H_i(\widetilde{C}_\bullet(K))$. Then $C_\bullet(K) = {}^{\widetilde{C}_\bullet(K)}\!\big/_{\langle e_\varnothing \rangle}$ is a chain complex, and $H_i(K) = H_i(C_\bullet(K))$ is the ith homology group of K.

Example. Let $K = \{e_0, e_1, \dots, e_r, e_\varnothing\}$, so $|K|$ is a collection of $r+1$ disjoint points. In this case, $\widetilde{C}_i(K) = 0$ for $i > 0$. $\widetilde{C}_0(K) = \langle e_0, \dots, e_r \rangle$ and $d(e_i) = \varnothing$. $\widetilde{C}_{-1}(K) = \langle e_\varnothing \rangle$. Hence $\ker d_0 = \langle e_1 - e_0, \dots, e_r - e_0 \rangle$ and $\operatorname{Im} d_1 = 0$, so $H_0(\widetilde{C}_\bullet(K)) = \mathbb{Z}^r$, and $H_{-1}(\widetilde{C}_\bullet(K)) = 0$. Note that $H_0(C_\bullet(K)) = \mathbb{Z}^{r+1} = \langle e_0, \dots, e_r \rangle$.

Example. Recall that any Euclidean simplicial complex is realised by an abstract simplicial complex, but we have choice in the labelling of the vertices. Let T_n be the boundary of a convex n-gon in $\mathbb{R}^2$. Then the abstract simplicial complex

$$K' = \{e_\varnothing, e_0, \dots, e_{n-1}, e_{01}, e_{12}, \dots, e_{(n-2)(n-1)}, e_{(n-1)0}\}$$

in Δ^{n-1} realises T_n. Then

$$\begin{aligned} C_1(K') &= \langle e_{01}, e_{12}, \dots, e_{(n-2)(n-1)}, e_{(n-1)0} \rangle \\ C_0(K') &= \langle e_0, \dots, e_{n-1} \rangle \end{aligned}$$

We have $d(e_{i(i+1)}) = e_{i+1} - e_i$, so $\ker d_1 = \langle x \rangle$ where

$$x = e_{01} + e_{12} + \dots + e_{(n-2)(n-1)} - e_{0(n-1)}$$

Note that $\operatorname{Im} d_1 = \operatorname{span}\{e_{i+1} - e_i\}$. Hence

$$\begin{aligned} H_1(K') &= {}^{\ker d_1}\!\big/_{\operatorname{Im} d_2} = {}^{\langle x \rangle}\!\big/_{0} \simeq \mathbb{Z} \\ H_0(K') &= {}^{\ker d_0}\!\big/_{\operatorname{Im} d_1} = {}^{\langle e_0, \dots, e_{n-1} \rangle}\!\big/_{\operatorname{span}\{e_1 - e_0, \dots, e_{n-1} - e_{n-2}\}} \simeq \mathbb{Z} \end{aligned}$$

Note that this result does not depend on the choice of n, and $|T_n| \simeq S^1$ also does not depend on n. In fact, $H_\bullet(K)$ depends only on $|K|$.

7.3. Chain maps

Definition. Let $(C_\bullet, d)$ and $(C'_\bullet, d')$ be chain complexes. A *chain map* $f : C_\bullet \to C'_\bullet$ is

(i) for each i, a homomorphism $f_i : C_i \to C'_i$, such that

(ii) $f_{i-1} \circ d_i = d'_i \circ f_i$.

Remark. We can interpret f as $\bigoplus_i f_i : C_\bullet \to C'_\bullet$, given by a block matrix

$$\begin{pmatrix} f_n & & \\ & f_{n-1} & \\ & & \ddots \end{pmatrix}$$

Then part (ii) of the definition is equivalent to the statement $d' f = f d$.

If $x \in \ker d$, we write $[x] \in H_\bullet(C)$ for its image under the map $\ker d \to {}^{\ker d}\!/\!_{\operatorname{Im} d}$.

Remark. $f(\ker d) \subseteq \ker d'$ because if $dx = 0$, we have $d'(f(x)) = f(d(x)) = f(0) = 0$. $f(\operatorname{Im} d) \subseteq \operatorname{Im} d'$, because if $x = dy$, we have $f(x) = f(d(y)) = d'(f(y))$. So f descends to a well-defined homomorphism $f_\star : {}^{\ker d}\!/\!_{\operatorname{Im} d} \to {}^{\ker d'}\!/\!_{\operatorname{Im} d'}$ such that $f_\star([x]) = [f(x)]$. So $f_\star : H_\bullet(C) \to H_\bullet(C')$. This is called the map *induced by f*.

Remark. The composition of two chain maps is a chain map, and $(f \circ g)_\star = f_\star \circ g_\star$.

Let K be an abstract simplicial complex in Δ^n, and L be an abstract simplicial complex in Δ^m. Let $f : K \to L$ be a simplicial map, so it is determined by $\hat{f} : \{0, \dots, n\} \to \{0, \dots, m\}$. We wish to define a chain map $f_\sharp : C_\bullet(K) \to C_\bullet(L)$, which will induce $f_\star : H_\bullet(K) \to H_\bullet(L)$. Perhaps the most obvious guess would be to define $f_\sharp(e_I) = f(e_I) = e_{\hat{f}(I)}$. This is not the correct definition.

First, consider $f : \mathbf{\Delta}^1 \to \mathbf{\Delta}^1$ given by $e_0 \mapsto e_0, e_1 \mapsto e_0$. Then $f(e_{01}) = e_0$, but $e_{01} \in C_1(\mathbf{\Delta}^1)$ and $e_0 \in C_0(\mathbf{\Delta}^1)$. So f does not preserve grading, and hence cannot be a chain map.

Consider also $f : \mathbf{\Delta}^1 \to \mathbf{\Delta}^1$ given by $e_0 \mapsto e_1$ and $e_1 \mapsto e_0$. Now, $f(e_{01}) = e_{01}, f(e_0) = e_1, f(e_1) = e_0$, so $df(e_{01}) = d(e_{01}) = e_1 - e_0$ but $fd(e_{01}) = f(e_1 - e_0) = e_0 - e_1$.

The solution to both problems is to change our perspective on the indices I. Until now, we have defined $I \subseteq \{0, \dots, n\}$ and written $I = i_0 i_1 \dots i_k$ where $i_0 < \dots < i_k$. Instead, we will allow $I \in \{0, \dots, n\}^{k+1}$, so $I = (i_0, i_1, \dots, i_k) = i_0 i_1 \dots i_k$ with no restriction on order. For instance, e_{00}, e_{10} are permitted.

We impose relations on the set of all such I to form an abelian group generated by equivalence classes of the $\{0, \dots, n\}^{k+1}$. We will define that $e_I = -e_{I'}$ when I, I' are related by switching two indices; so $e_{102} = -e_{012} = e_{210}$. If e_I contains a repetition, we require $e_I = 0$.

More concretely, if $I \in \{0, \dots, n\}^{k+1}$, let I' be the unique ordered permutation of I if I has no repetitions. Then $e_I = (-1)^{S(I)} e_{I'}$ if I has no repetitions, and $e_I = 0$ if I has a repetition, where $(-1)^{S(I)}$ is the sign of the permutation $\sigma \in S^{k+1}$ mapping I to I'. If we draw I and I' in order as a bipartite planar graph, connected by matching labels, $S(I)$ is the number of crossings.

Lemma. Let $i_j \in I$, and suppose i_j is in position $i_{j'}$ in I'. Then $S(I) - S(I_{\hat{j}}) \equiv j - j' \bmod 2$.

Proposition. Let $I \in \{0, \dots, n\}^{k+1}$. Then $d(e_I) = \sum_{j=0}^{k} (-1)^j e_{I_{\hat{j}}}$, where $I_{\hat{j}}$ is obtained from I by omitting the jth entry in the tuple I.

We have already defined d for ordered sequences of indices; this proposition states that this formula holds for all sequences of indices.

Proof.

$$\sum_{j=0}^{k}(-1)^k e_{I_j} = \sum_{j=0}^{k}(-1)^j(-1)^{S(I_j)}e_{I'_j} = \sum_{j=0}^{k}(-1)^{j'}(-1)^{S(I)}e_{(I')_j} = (-1)^{S(I)}d(e_{I'}) = d(e_I)$$

□

Example. $d(e_{210}) = (-1)^0 e_{10} + (-1)^1 e_{20} + (-1)^2 e_{21} = -e_{01} + e_{02} - e_{12} = d(-e_{012})$, where by definition, $e_{210} = -e_{012}$ so $d(e_{210}) = -d(e_{012})$.

Definition. Let $f : K \to L$ be a simplicial map. Then $f_\sharp : C_k(K) \to C_k(L)$ is defined by $f_\sharp(e_I) = e_{\hat{f}(I)}$ where if $I = (i_0, \dots, i_k)$ we define $\hat{f}(I) = (\hat{f}(i_0), \dots, \hat{f}(i_k))$.

This definition of $f_\sharp$ preserves grading.

Proposition. $f_\sharp$ is a chain map.

Proof.

$$d(f_\sharp(e_I)) = d(e_{\hat{f}(I)}) = \sum_{j=0}^{k}(-1)^j e_{(\hat{f}(I))_j} = f_\sharp\left(\sum_{j=0}^{k}(-1)^j e_{I_j}\right) = f_\sharp(d(e_I))$$

□

Example. Let $f : \mathbf{\Delta}^1 \to \mathbf{\Delta}^1$ be the simplicial map defined by $f(e_0) = e_0$ and $f(e_1) = e_0$. Then $f_\sharp(e_{01}) = e_{00} = 0$.

Now let $f(e_0) = e_1$ and $f(e_1) = e_0$. Then $f_\sharp(e_{01}) = e_{10} = -e_{01}, f_\sharp(e_0) = e_1, f_\sharp(e_1) = e_0$. So $d(f_\sharp(e_{01})) = -d(e_{01}) = e_0 - e_1 = f(d(e_{01}))$.

7.4. Chain homotopies

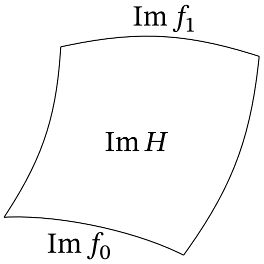
Definition. Let $f_0, f_1 : (C, d) \to (C', d')$ be chain maps. Then f_0 is *chain homotopic* to f_1, written $f_0 \sim f_1$, if there are

(i) homomorphisms $h_i : C_i \to C'_{i+1}$, where we write $h = \bigoplus_i h_i$, satisfying

(ii) $d'h + hd = f_0 - f_1$.

In this case, we say h is the *chain homotopy*.

Example. Suppose $f_0, f_1 : X \to Y$ are homotopic maps via H. Suppose $X = \Delta^n$.



Here,

$$\partial(H(\Delta^n)) = H(\partial\Delta^n) \cup f_1(\Delta^n) \cup f_0(\Delta^n) \implies \partial H + H\partial = f_1 + f_0$$

without considering signs.

Lemma. If $f_1 \sim f_0$, then $f_{1\star} = f_{0\star} : H_\bullet(C) \to H_\bullet(C')$.

Proof. Let $[x] \in H_\bullet(C)$. Then $dx = 0$. So

$$f_{1\star}[x] - f_{0\star}[x] = [(f_1 - f_0)x] = [(d'h + hd)x] = [d'h(x)] = 0$$

since $d'h(x) \in \operatorname{Im} d'$. □

Definition. We say a chain complex (C, d) is *contractible* if $\mathrm{id}_C \sim 0_C$, where 0_C is the zero map.

Lemma. Let (C, d) be contractible. Then $H_\bullet(C) = 0$.

Proof. Let $[x] \in H_\bullet(C)$. Then $[x] = \mathrm{id}_\star[x] = 0_\star[x] = [0]$. So $H_k(C)$ is the trivial group for each k. □

Definition. Let K be an abstract simplicial complex in Δ^n. Let $e_0 \notin K$. Then the *cone* is $C_{e_0}(K) = K \cup \{e_{0I} \mid e_I \in K\}$.

Remark. $C_{e_0}(K)$ is an abstract simplicial complex. If K' is a realisation of K, where $e_0 \notin K$ and K' is independent of p, then $C_p(K')$ is a realisation of $C_{e_0}(K)$.

Example. Consider $\hat{\boldsymbol{\Delta}}^n = \left\{e_I \in \boldsymbol{\Delta}^{n+1} \mid 0 \notin I\right\} \simeq \boldsymbol{\Delta}^n$. Then $C_{e_0}(\hat{\boldsymbol{\Delta}}^n) = \boldsymbol{\Delta}^{n+1}$.

Proposition. $\widetilde{C}_\bullet(C_{e_0}(K))$ is contractible.

Proof. Define $h : \widetilde{C}_k(C_{e_0}(K)) \to \widetilde{C}_{k+1}(C_{e_0}(K))$ by $h(e_I) = e_{0I}$. Note that if $0 \in I$, then $e_{0I} = 0$.

If $0 \in I$ then $dh(e_I) = 0$, and $hd(e_I) = h\left(\sum_{j=0}^k (-1)^j e_{I_j}\right) = h(e_{I \setminus \{0\}} + \sum e_{I'})$ where $0 \in I'$. Then $hd(e_I) = e_I + 0 = e_I$. Otherwise, if $0 \notin I$, then $dh(e_I) = d(e_{0I}) = e_I + \sum_{j=0}^k (-1)^{k+1} e_{0I_j} = e_I - h(de_I)$. In either case, $dh + hd = \mathrm{id}$. □

Corollary. $\widetilde{H}_\bullet(C_{e_0}(K)) = 0$. In particular,

$$H_i(C_{e_0}(K)) = \begin{cases} \mathbb{Z} & i = 0 \\ 0 & i \neq 0 \end{cases}$$

Proof. Let $\widetilde{C}_\bullet(C_{e_0}(K)) = (\widetilde{C}, \widetilde{d})$, and $C_\bullet(C_{e_0}(K)) = (C, d)$. The first part follows from the previous result. For the second part, note that $\widetilde{H}_{-1}(C_{e_0}(K)) = 0$, so $\widetilde{d}_0 : \widetilde{C}_0 \to \widetilde{C}_{-1} = \langle e_\varnothing \rangle \simeq \mathbb{Z}$ is surjective. So $\mathbb{Z} \simeq \operatorname{Im} \widetilde{d}_0 \simeq {}^{\widetilde{C}_0}\!/\!_{\ker \widetilde{d}_0} \simeq {}^{\widetilde{C}_0}\!/\!_{\operatorname{Im} \widetilde{d}_1}$ since $\widetilde{H}_0(C) = 0$. But ${}^{\widetilde{C}_0}\!/\!_{\operatorname{Im} \widetilde{d}_1} \simeq$

${}^{C_0}\!/\!{}_{\operatorname{Im} d_1} = {}^{\ker d_0}\!/\!{}_{\operatorname{Im} d_1} = H_0(C_{e_0}(K))$. For $i \geq 0$, note that $\ker \widetilde{d}_i = \ker d_i$ and $\operatorname{Im} \widetilde{d}_{i+1} = \operatorname{Im} d_{i+1}$. Hence $H_i(\widetilde{C}) = H_i(C)$ for $i > 0$. □

Definition. Let $\mathbb{S}^n = \mathbf{\Delta}^n \setminus e_{0\dots n}$. Then

$$H_i(\mathbb{S}^n) = \begin{cases} \mathbb{Z} & i = 0, n \\ 0 & \text{otherwise} \end{cases}$$

Proof. Similar to the previous proof, but now we remove the 'top' generator instead of the 'bottom' one. □

Alternatively, we can prove this fact using the results from the next subsection.

7.5. Exact sequences

Definition. Let A_k be a sequence of abelian groups for $k \in \mathbb{Z}$, and $f_k : A_k \to A_{k-1}$ be homomorphisms. We say that the sequence is *exact* at A_k if $\ker f_k = \operatorname{Im} f_{k+1}$. If it is exact at all A_k, we say the sequence is exact.

$$\begin{tikzcd} \cdots \arrow[r, "f_{k+2}"] & A_{k+1} \arrow[r, "f_{k+1}"] & A_k \arrow[r, "f_k"] & A_{k-1} \arrow[r, "f_{k-1}"] & \cdots \end{tikzcd}$$

Example.

$$\begin{tikzcd} 0 \arrow[r] & A \arrow[r, "f"] & B \end{tikzcd}$$

is exact at A if and only if f is injective.

$$\begin{tikzcd} B \arrow[r, "g"] & C \arrow[r] & 0 \end{tikzcd}$$

is exact at C if and only if g is surjective.

$$\begin{tikzcd} 0 \arrow[r] & A \arrow[r, "f"] & B \arrow[r, "g"] & C \arrow[r] & 0 \end{tikzcd}$$

is exact if and only if f is injective, g is surjective, and $g : {}^{B}\!/\!{}_{\operatorname{Im} f} \to C$ is an isomorphism, so $C \simeq {}^{B}\!/\!{}_{\operatorname{Im} f}$. An exact sequence of the form

$$\begin{tikzcd} 0 \arrow[r] & A \arrow[r, "f"] & B \arrow[r, "g"] & C \arrow[r] & 0 \end{tikzcd}$$

is called a *short exact sequence.*

Definition. Let $g : B \to C$. Then the *cokernel* of g is $\operatorname{coker} g = {}^{C}\!/\!{}_{\operatorname{Im} g}$.

In general, a sequence is exact if and only if

$$\begin{tikzcd} 0 \arrow[r] & \operatorname{coker} f_{k+1} \arrow[r, "f_{k+1}"] & A_k \arrow[r, "f_k"] & \ker f_{k-1} \arrow[r] & 0 \end{tikzcd}$$

is a short exact sequence for every k.

Definition. A short exact sequence of chain complexes is a short exact sequence

$$\begin{tikzcd}
0 \arrow[r] & A_\bullet \arrow[r, "f"] & B_\bullet \arrow[r, "g"] & C_\bullet \arrow[r] & 0
\end{tikzcd}$$

where $A_\bullet, B_\bullet, C_\bullet$ are chain complexes, and f, g are chain maps.

Equivalently, we have the diagram

$$\begin{tikzcd}
& \vdots \arrow[d, "d_A"] & \vdots \arrow[d, "d_B"] & \vdots \arrow[d, "d_C"] & \\
0 \arrow[r] & A_k \arrow[r, "f"] \arrow[d, "d_A"] & B_k \arrow[r, "g"] \arrow[d, "d_B"] & C_k \arrow[r] \arrow[d, "d_C"] & 0 \\
0 \arrow[r] & A_{k-1} \arrow[r, "f"] \arrow[d, "d_A"] & B_{k-1} \arrow[r, "g"] \arrow[d, "d_B"] & C_{k-1} \arrow[r] \arrow[d, "d_C"] & 0 \\
0 \arrow[r] & A_{k-2} \arrow[r, "f"] \arrow[d, "d_A"] & B_{k-2} \arrow[r, "g"] \arrow[d, "d_B"] & C_{k-2} \arrow[r] \arrow[d, "d_C"] & 0 \\
& \vdots & \vdots & \vdots &
\end{tikzcd}$$

where all squares commute since f, g are chain maps, and all rows are exact.

Lemma (snake lemma). Let $0 \longrightarrow A_\bullet \xrightarrow{f} B_\bullet \xrightarrow{g} C_\bullet \longrightarrow 0$ be a short exact sequence of chain complexes. Then there is an exact sequence

$$\begin{tikzcd}
H_k(A) \arrow[r, "f_\star"] & H_k(B) \arrow[r, "g_\star"] & H_k(C) \arrow[dll, rounded corners, "\partial_k"] \\
H_{k-1}(A) \arrow[r, "f_\star"] & H_{k-1}(B) \arrow[r, "g_\star"] & H_{k-1}(C)
\end{tikzcd}$$

The homomorphism ∂_k is called the *connecting homomorphism*. Since this exists for all k, this gives a long exact sequence of homology groups.

Proof. Let $[c] \in H_k(C)$, so $dc = 0$. Then,

(i) g is surjective, so we can choose $b \in B_k$ such that $g(b) \in c$.

(ii) $g(db) = dg(b) = dc = 0$, so $db \in \ker g$. Since the sequence is exact at B, we have $db = f(a)$ for some $a \in A_{k-1}$.

(iii) $f(da) = d(fa) = d^2(b) = 0$. Since f is injective, $da = 0$.

We then define $\partial_k[c] = [a] \in H_{k-1}(A)$. To visualise the above argument, the following diagrams can be overlaid; the first diagram shows the groups, and the second diagram shows the corresponding elements.

$$\begin{tikzcd}
0 \arrow[r] & A_k \arrow[r, "f"] \arrow[d, "d_A"] & B_k \arrow[r, "g"] \arrow[d, "d_B"] & C_k \arrow[r] \arrow[d, "d_C"] \arrow[dll, dashed, "\partial_k"] & 0 \\
0 \arrow[r] & A_{k-1} \arrow[r, "f"] & B_{k-1} \arrow[r, "g"] & C_{k-1} \arrow[r] & 0
\end{tikzcd}$$

$$\begin{tikzcd}
 & b \arrow[r, mapsto, "g"] \arrow[d, "d_B"] & c \arrow[dll, dashed, "\partial_k"] \\
a \arrow[r, mapsto, "f"] & db &
\end{tikzcd}$$

This definition does not depend on any choices that we made; for example, $[c] = [c']$ implies $\partial_k[c] = \partial_k[c']$.

(i) If $g(b') = c$, then $g(b - b') = 0$. By exactness, $b - b' = f(\alpha)$. Then $db - db' = f(d(\alpha))$. Let $f(a) = db$ and $f(a') = db'$. So $a - a' = d\alpha$, so $[a] = [a']$.

(ii) Suppose $[c] = [c']$. Then $c - c' = d\gamma$ for $\gamma \in C_{k+1}$. g is surjective, so let $\gamma = g(\beta)$. Then $b - b' = d\beta$, so $db = db'$. Since $a = a'$, we have $[a] = [a']$.

We need to check exactness. We will show that ker $\subseteq$ Im in each case, the other direction is left as an exercise.

(i) Consider $H_k(C)$. If $\partial_k[c] = 0$, then $a = d\alpha$ for $\alpha \in A_k$. Then $d(f(\alpha)) = f(d\alpha) = f(a) = db$. So $d(b - f(\alpha)) = 0$, giving $[b - f(\alpha)] \in H_k(B)$. Then $g_\star[b - f(\alpha)] = [g(b) - g(f(\alpha))] = [g(b)] = [c]$ by exactness. So $[c] \in \operatorname{Im} g_\star$ as required.

(ii) Consider $H_k(B)$. If $g_\star[b] = 0$, then $g(b) = d\gamma$ for some $\gamma \in C_{k+1}$. g is surjective, so $\gamma = g(\beta)$ for $\beta \in B_{k+1}$. Then $g(b - d\beta) = c - dg(\beta) = c - c = 0$, so $b - d\beta = f(\alpha)$ for $\alpha \in A_k$. So $f(d\alpha) = df(\alpha) = db - d^2\beta = 0$. Hence $[b] = [b - d\beta] = f_\star[\alpha]$. So $[\beta] \in \operatorname{Im} f_\star$.

(iii) Consider $H_{k-1}(A)$. If $f_\star[a] = 0$, then $f(a) = db$ for some b in B_{k-1}. Then $[a] = \partial_k[g(b)]$ since $dg(b) = g(db) = g(f(a)) = 0$. So $[a] \in \operatorname{Im} \partial_k$.

□

Example. Let $B = C_\bullet(\mathbf{\Delta}^n)$, and $A = C_\bullet(\mathbb{S}^{n-1})$. Let C be defined by

$$C_k = \begin{cases} \langle e_{0\dots n} \rangle & k = n \\ 0 & \text{otherwise} \end{cases}$$

Note that

$$H_k(C) = \begin{cases} \mathbb{Z} & k = n \\ 0 & \text{otherwise} \end{cases}$$

Let $n > 1$. Then we have a short exact sequence $0 \longrightarrow \mathbb{S}^{n-1} \xrightarrow{f} \mathbf{\Delta}^n \xrightarrow{g} C \longrightarrow 0$ and hence we have

$$\begin{tikzcd}
H_k(\mathbb{S}^{n-1}) \arrow[r, "f_\star"] & H_k(\mathbf{\Delta}^n) \arrow[r, "g_\star"] & H_k(C) \arrow[dll, rounded corners, to path={ -- ([xshift=2ex]\tikztostart.east) |- ([yshift=-1ex]$(\tikztostart)!0.5!(\tikztotarget)$) \tikztonodes -| ([xshift=-2ex]\tikztotarget.west) -- (\tikztotarget)}, "\partial_k"] \\
H_{k-1}(\mathbb{S}^{n-1}) \arrow[r, "f_\star"] & H_{k-1}(\mathbf{\Delta}^n) \arrow[r, "g_\star"] & H_{k-1}(C)
\end{tikzcd}$$

Now, letting $k = n$, we can therefore find the exact sequence

$$\begin{tikzcd}
H_n(\mathbb{S}^{n-1}) \arrow[r] & 0 \arrow[r] & \mathbb{Z} \arrow[dll, rounded corners, to path={ -- ([xshift=2ex]\tikztostart.east) |- ([yshift=-1ex]$(\tikztostart)!0.5!(\tikztotarget)$) \tikztonodes -| ([xshift=-2ex]\tikztotarget.west) -- (\tikztotarget)}, "\partial_k"] \\
H_{n-1}(\mathbb{S}^{n-1}) \arrow[r] & 0 \arrow[r] & 0
\end{tikzcd}$$

By exactness at $\mathbb{Z}$ and $H_{n-1}(\mathbb{S}^{n-1})$, ∂_k is an isomorphism. Hence $H_{n-1}(\mathbb{S}^{n-1}) = \mathbb{Z}$.

7.6. Mayer–Vietoris sequence

Let K_1, K_2 be abstract simplicial complexes in Δ^n. Then $K_1 \cap K_2$ and $K_1 \cup K_2$ are also abstract simplicial complexes in Δ^n. We have the following commutative square of simplicial maps given by inclusion.

$$\begin{tikzcd}
 & K_1 \arrow[dr, "j_1"] & \\
K_1 \cap K_2 \arrow[ur, "i_1"] \arrow[dr, "i_2"'] & & K_1 \cup K_2 \\
 & K_2 \arrow[ur, "j_2"'] &
\end{tikzcd}$$

This induces a commutative square of chain maps as shown.

$$\begin{tikzcd}
 & C_\bullet(K_1) \arrow[dr, "j_{1\sharp}"] & \\
C_\bullet(K_1 \cap K_2) \arrow[ur, "i_{1\sharp}"] \arrow[dr, "i_{2\sharp}"'] & & C_\bullet(K_1 \cup K_2) \\
 & C_\bullet(K_2) \arrow[ur, "j_{2\sharp}"'] &
\end{tikzcd}$$

Proposition. Let K_1, K_2 be abstract simplicial complexes in Δ^n. Then the sequence

$$\begin{tikzcd} 0 \arrow[r] & C_\bullet(K_1 \cap K_2) \arrow[r, "i"] & C_\bullet(K_1) \oplus C_\bullet(K_2) \arrow[r, "j"] & C_\bullet(K_1 \cup K_2) \arrow[r] & 0 \end{tikzcd}$$

is a short exact sequence of chain complexes, where

$$i = \begin{pmatrix} i_{1\sharp} \\ i_{2\sharp} \end{pmatrix}; \quad j = \begin{pmatrix} j_{1\sharp} & -j_{2\sharp} \end{pmatrix}$$

Proof. We must check exactness at each location. $i_{1\sharp}$ is injective, so i is injective.

If $j((a,b)) = 0$, then $j_{1\sharp}(a) = j_{2\sharp}(b)$, so $a = b \in C_\bullet(K_1) \cap C_\bullet(K_2) = C_\bullet(K_1 \cap K_2)$. Hence $(a,b) = i(a)$, so $\ker j \subseteq \operatorname{Im} i$. For the other direction, $gf(a) = (j_{1\sharp} \circ i_{1\sharp})(a) - (j_{2\sharp} \circ i_{2\sharp}(a)) = 0$ since the square of inclusion maps commutes. So $\operatorname{Im} i \subseteq \ker j$, so the sequence is exact at $C_\bullet(K_1) \oplus C_\bullet(K_2)$.

Let $e_I \in K_1 \cup K_2$. Then $e_I \in K_1$ or $e_I \in K_2$. If $e_I \in K_1$ then $e_I = j((e_I, 0))$. If $e_I \in K_2$ then $e_I = j((0, -e_I))$. So $e_I \in \operatorname{Im} j$ in either case. Since the e_I form a free basis, j is surjective as required. □

Theorem (Mayer–Vietoris sequence). Let K_1, K_2 be abstract simplicial complexes in Δ^n. Then there is a long exact sequence

$$\begin{tikzcd} H_k(K_1 \cap K_2) \arrow[r, "i_\star"] & H_k(K_1) \oplus H_k(K_2) \arrow[r, "j_\star"] & H_k(K_1 \cup K_2) \arrow[dll, rounded corners, "\partial_k"] \\ H_{k-1}(K_1 \cap K_2) \arrow[r, "i_\star"] & H_{k-1}(K_1) \oplus H_{k-1}(K_2) \arrow[r, "j_\star"] & H_{k-1}(K_1 \cup K_2) \end{tikzcd}$$

Proof. Follows from the above theorem and the snake lemma. □

Example. Let K_1, K_2 be abstract simplicial complexes in Δ^n, Δ^m. Then let $K_1 \amalg K_2 \subset \Delta^{n+m+1}$ be the abstract simplicial complex where the vertices of Δ^{n+m+1} are $e_0, \dots, e_n, e_0', \dots, e_m'$, and we embed K_1 and K_2 into $K_1 \amalg K_2$ in the natural way. More precisely, $e_I \in K_1$ gives $e_I \in K_1 \amalg K_2$, and $e_I \in K_2$ gives $e_J' \in K_1 \amalg K_2$. Then $|K_1 \amalg K_2| = |K_1| \amalg |K_2|$. $K_1 \amalg K_2 = K_1 \cup K_2'$ where K_1, K_2' are disjoint abstract simplicial complexes in Δ^{n+m+1}, so $K_1 \cap K_2' = \{e_\varnothing\}$. The Mayer–Vietoris sequence gives

$$\begin{tikzcd} H_k(\{e_\varnothing\}) \arrow[r, "i_\star"] & H_k(K_1) \oplus H_k(K_2) \arrow[r, "j_\star"] & H_k(K_1 \amalg K_2) \arrow[dll, rounded corners, "\partial_k"] \\ H_{k-1}(\{e_\varnothing\}) \arrow[r, "i_\star"] & H_{k-1}(K_1) \oplus H_{k-1}(K_2) \arrow[r, "j_\star"] & H_{k-1}(K_1 \amalg K_2) \end{tikzcd}$$

Note that $H_k(\{e_\varnothing\}) = 0$. Hence, the sequence

$$\begin{tikzcd} 0 \arrow[r] & H_k(K_1) \oplus H_k(K_2) \arrow[r] & H_k(K_1 \amalg K_2) \arrow[r] & 0 \end{tikzcd}$$

is exact. So $H_k(K_1) \oplus H_k(K_2) \simeq H_k(K_1 \amalg K_2)$.

7.7. Homology of triangulable spaces

Theorem. Let $f_0, f_1 : K \to L$ be simplicial approximations to a continuous map $F : |K| \to |L|$. Then $f_{0\sharp} \sim f_{1\sharp}$, so $f_{0\star} = f_{1\star}$.

Theorem. There is an isomorphism $\nu_K : H_\bullet(BK) \to H_\bullet(K)$ such that $\nu_K = f_\star$ where $f : BK \to K$ is any simplicial approximation to the identity map on $|K|$.

Definition. Let $F : |K| \to |L|$ be continuous. By the simplicial approximation theorem, there exists $f : B^r \to L$ that is a simplicial approximation to F. Define $F_\star : H_\bullet(K) \to H_\bullet(L)$ by $F_\star = f_\star \circ \nu_{K,r}^{-1}$.

Theorem. $F_\star$ is well-defined, so does not depend on the choice of f. $(\mathrm{id}_K)_\star = \mathrm{id}_{H_\bullet(K)}$. Further, $(F \circ G)_\star = F_\star \circ G_\star$.

Theorem. Let $F_0, F_1 : |K| \to |L|$ be continuous with $F_0 \sim F_1$. Then $F_{0\star} = F_{1\star}$.

Proposition. Let $|K| \sim |L|$. Then $H_\bullet(K) \simeq H_\bullet(L)$.

Proof. Let $F : |K| \to |L|$ and $G : |L| \to |K|$ be functions such that $F \circ G \sim \mathrm{id}_{|L|}$ and $G \circ F \sim \mathrm{id}_{|K|}$. Then $F_\star \circ G_\star = \mathrm{id}_{H_\bullet(L)}$ and $G_\star \circ F_\star = \mathrm{id}_{H_\bullet(K)}$ by functoriality. Hence $F_\star$ and $G_\star$ are inverse isomorphisms of groups. □

Definition. A space X is *triangulable* if there exists an abstract simplicial complex K with $|K| \simeq X$.

Remark. The above proposition implies that if X is triangulable, there is a well-defined homology group $H_\bullet(X) = H_\bullet(K)$ where K is any abstract simplicial complex with polyhedron $|K| \simeq X$. Not all topological spaces are homotopy equivalent to a triangulable space. One example is $\bigvee_{i=1}^{\infty} S^1$.

Proposition. Let $|K|$ be path-connected. Then $H_0(K) \simeq \mathbb{Z}$.

Proof. $C_0(K)$ is generated by the vertices e_i of K. Consider $F_i : \Delta^0 \to |K|$ mapping $e_0 \in \Delta^0$ to $e_i \in K$. Then $H_\star(\Delta^0) = \mathbb{Z} = \langle [e_0] \rangle$, and $F_\star([e_0]) = [e_i]$. Since K is path-connected, $F_i \sim F_j$. So $[e_i] = F_{i\star}([e_0]) = F_{j\star}([e_0]) = [e_j]$. Hence all $[e_i]$ are equal. The $[e_i]$ are not boundaries, so $H_0(K)$ is not trivial. □

Corollary. $H_0(K) = \mathbb{Z}^k$ where k is the number of path-connected components of $|K|$.

Proof. $|K|$ is a disjoint union of the k path-connected components of $|K|$, so $H_0(K)$ is the direct sum of the homology groups of these components. □

We know $S^n \simeq |\mathbb{S}^n|$, so

$$H_k(S^n) = H_k(\mathbb{S}^n) = \begin{cases} \mathbb{Z} & k = 0, n \\ 0 & \text{otherwise} \end{cases}$$

Hence $S^n \sim S^m$ implies $n = m$.

Corollary. $\mathbb{R}^n \simeq \mathbb{R}^m$ implies $n = m$.

Proof. Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a homeomorphism. Then $S^{n-1} \sim \mathbb{R}^n \setminus \{0\} \simeq \mathbb{R}^m \setminus \{f(0)\} \sim S^{m-1}$. So $S^{n-1} \sim S^{m-1}$, giving $n = m$. □

Corollary. There is no retraction $r : D^n \to S^{n-1}$.

Proof. We suppose $n > 0$. Let $j : S^{n-1} \to D^n$ be the inclusion. r is a retraction if and only if $r \circ j = \mathrm{id}_{S^{n-1}}$. This gives $(r \circ j)_\star = \mathrm{id}_{H_\cdot(S^{n-1})}$. Note that $H_{n-1}(D^n) = H_{n-1}(\boldsymbol{\Delta}^n) = 0$, and $H_{n-1}(S^{n-1}) = \mathbb{Z}$. If r is a retraction, then $r_\star$ and $j_\star$ are inverse homomorphisms of groups, but $\mathbb{Z}$ is not isomorphic to 0. So r is not a retraction. □

Theorem (Brouwer fixed point theorem)**.** Let $F : D^n \to D^n$ be a continuous function. Then F has a fixed point.

Remark. This is a generalisation of the intermediate value theorem for high dimensions.

Proof. Suppose there is no fixed point. Then, we define $G : D^n \to S^{n-1}$ by letting $G(x), x, F(x)$ lie in this order on a straight line in D^n. If G is a well-defined continuous map, it is a retraction, contradicting the previous result.

Let $p \in D^n$ and $v \in S^{n-1}$. Let $R_{p,v} = \{p + tv \mid t \geq 0\}$. If $p{+}tv \in S^{n-1}$, then $\langle p + tv, p + tv\rangle = 1$, so $\langle p, p\rangle + 2t\langle v, p\rangle + t^2 = 1$. Hence

$$t = -\langle p, v\rangle \pm \sqrt{\langle p, v\rangle^2 + 1 - \langle p, p\rangle}$$

We define

$$\tau(p, v) = \max\left(-\langle p, v\rangle \pm \sqrt{\langle p, v\rangle^2 + 1 - \langle p, p\rangle}\right)$$

This is a continuous function. Now, we define $P(p, v) = p + \tau(p, v)v$, which is the intersection of $R_{p,v}$ with S^{n-1}, which is also continuous. So

$$G(x) = P\left(F(x), \frac{x - F(x)}{\|x - F(x)\|}\right)$$

is well-defined and continuous. □

7.8. Homology of orientable surfaces

We can often compute homology groups only using the Mayer–Vietoris sequence and functoriality properties.

Example. Consider the torus T^2. We can write a triangulation K of T^2 as $K_1 \cup K_2$, with $|K_i| \simeq S^1 \times I$, and $|K_1 \cap K_2| \simeq S^1_1 \amalg S^1_2$. Note that the inclusion $\iota_{j,i} : S^1_j \hookrightarrow |K_i|$ is a homotopy equivalence, and $\iota_{1,i} \sim \iota_{2,i}$. Then the Mayer–Vietoris sequence gives

$$\begin{tikzcd}
 & H_2(K_1) \oplus H_2(K_2) \arrow[r] & H_2(K) \arrow[dll, rounded corners] & \\
H_1(K_1 \cap K_2) \arrow[r, "\alpha_1"'] & H_1(K_1) \oplus H_1(K_2) \arrow[r] & H_1(K) \arrow[dll, rounded corners] & \\
H_0(K_1 \cap K_2) \arrow[r, "\alpha_0"'] & H_0(K_1) \oplus H_0(K_2) \arrow[r] & H_0(K) \arrow[r] & 0
\end{tikzcd}$$

giving

$$\begin{tikzcd}
 & 0 \arrow[r] & H_2(K) \arrow[dll, rounded corners] & \\
\mathbb{Z} \oplus \mathbb{Z} \arrow[r, "\alpha_1"'] & \mathbb{Z} \oplus \mathbb{Z} \arrow[r] & H_1(K) \arrow[dll, rounded corners] & \\
\mathbb{Z} \oplus \mathbb{Z} \arrow[r, "\alpha_0"'] & \mathbb{Z} \oplus \mathbb{Z} \arrow[r] & H_0(K) \arrow[r] & 0
\end{tikzcd}$$

Hence we have short exact sequences

$$\begin{tikzcd}
0 \arrow[r] & H_2(K) \arrow[r] & \ker \alpha_1 \arrow[r] & 0
\end{tikzcd}$$

$$\begin{tikzcd}
0 \arrow[r] & \operatorname{coker} \alpha_1 \arrow[r] & H_1(K) \arrow[r] & \ker \alpha_0 \arrow[r] & 0
\end{tikzcd}$$

$$\begin{tikzcd}
0 \arrow[r] & \operatorname{coker} \alpha_0 \arrow[r] & H_0(K) \arrow[r] & 0
\end{tikzcd}$$

The maps α_i are given by the matrix $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$. Therefore, $\ker \alpha_i \simeq \mathbb{Z}$ and $\operatorname{coker} \alpha_i \simeq \mathbb{Z}$. Hence $H_2(K) \simeq \mathbb{Z}$, $H_1(K) \simeq \mathbb{Z}^2$, and $H_0(K) \simeq \mathbb{Z}$.

$$H_k(T^2) = \begin{cases} \mathbb{Z} & k = 0, 2 \\ \mathbb{Z}^2 & k = 1 \\ 0 & \text{otherwise} \end{cases}$$

Proposition. Suppose that $0 \longrightarrow A \longrightarrow B \longrightarrow \mathbb{Z}^r \longrightarrow 0$ is exact. Then $B \simeq A \oplus \mathbb{Z}^r$.

Proof. By exactness, $\mathbb{Z}^r \simeq {}^{B}\!/_{A}$. The result then follows from the structure theorem for abelian groups. □

Example. Let L_1 be a triangulation of T^2, and let $L_{1,1}$ be $L_1 \setminus \{\sigma\}$ where σ is a 2-simplex. Then $\partial L_{1,1} \simeq \partial\sigma = \mathbb{S}^1$, and $|L_{1,1}| \sim S^1 \vee S^1$. We inductively define $L_g = L_{g-1,1} \cup_{\mathbb{S}^1} L_{1,1}$, and $L_{g,1} = L_g \setminus \sigma$ where σ is a 2-simplex. Then L_g is a triangulation of the compact surface of genus g. Note also that $L_{g,1} \simeq L_{g-1,1} \cup_{\sigma^1} L_{1,1}$ where σ^1 is an edge of $\mathbb{S}^1$. So $L_{g,1} \sim \bigvee_{i=1}^{2g} S^1$.

Proposition.

$$H_k(L_g) = \begin{cases} \mathbb{Z} & k = 0, 2 \\ \mathbb{Z}^{2g} & k = 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$H_k(L_{g,1}) = \begin{cases} \mathbb{Z} & k = 0 \\ \mathbb{Z}^{2g} & k = 1 \\ 0 & \text{otherwise} \end{cases}$$

Further, $\iota_{g\star} : H_1(\partial L_{g,1}) \to H_1(L_{g,1})$ is the zero map.

Proof. By induction, we show the result for $H_k(L_g)$ implies the result for $H_k(L_{g,1})$, and then $H_k(L_{g,1})$ gives $H_k(L_{g+1})$. The base case is $H_\bullet(T^2)$ which was shown above. For the first implication, we use the Mayer–Vietoris sequence. Note that $L_g = L_{g,1} \cup_{\partial L_{g,1}} \mathbf{\Delta}^2$. Then,

$$\begin{tikzcd}
 & H_2(L_{g,1}) \oplus H_2(\mathbf{\Delta}^2) \arrow[r] & H_2(L_g) \arrow[dll, rounded corners, "\partial_2" description] \\
H_1(\partial L_{g,1}) \arrow[r, "\iota_1"'] & H_1(L_{g,1}) \oplus H_1(\mathbf{\Delta}^2) \arrow[r] & H_1(L_g) \arrow[dll, rounded corners, "\partial_1" description] \\
H_0(\partial L_{g,1}) \arrow[r, "\iota_0"'] & H_0(L_{g,1}) \oplus H_0(\mathbf{\Delta}^2) \arrow[r] & H_0(L_g)
\end{tikzcd}$$

giving

$$\begin{tikzcd}
 & 0 \oplus 0 \arrow[r] & \mathbb{Z} \arrow[dll, rounded corners, "\partial_2" description] \\
\mathbb{Z} \arrow[r, "\iota_1"'] & H_1(L_{g,1}) \oplus 0 \arrow[r] & \mathbb{Z}^{2g} \arrow[dll, rounded corners, "\partial_1" description] \\
\mathbb{Z} \arrow[r, "\iota_0"'] & \mathbb{Z} \oplus \mathbb{Z} \arrow[r] & \mathbb{Z}
\end{tikzcd}$$

The bottom row of the Mayer–Vietoris sequence always has this form if $K_1, K_2, K_1 \cap K_2$ are connected. Note that since ι_0 is injective, the map before it is the zero map by exactness, so we can remove the bottom row and replace it with zero. We have that ∂_2 is injective, and $H_1(L_{g,1})$ is torsion-free, so ∂_2 is an isomorphism. Hence ι_1 is the zero map and j is an isomorphism. Since $0 = \iota_1 = \iota_{g\star} + \iota'_\star$, we have $\iota_{g\star} = 0$. Further, as j is an isomorphism, $H_1(L_{g,1}) \simeq H_1(L_g) = \mathbb{Z}^{2g}$ as required.

Now we show the result for $H_k(L_{g,1})$ implies the result for $H_k(L_{g+1})$. Note that $L_{g+1} = L_{g,1} \cup_{\partial L_{g,1}} L_{1,1}$. Hence,

$$\begin{tikzcd}
 & H_2(L_{g,1}) \oplus H_2(L_{1,1}) \arrow[r] & H_2(L_{g+1}) \arrow[dll] & \\
H_1(\partial L_{g,1}) \arrow[r, "\iota"'] & H_1(L_{g,1}) \oplus H_2(L_{1,1}) \arrow[r] & H_1(L_{g+1}) \arrow[r] & 0
\end{tikzcd}$$

so

$$\begin{tikzcd}
 & 0 \oplus 0 \arrow[r] & H_2(L_{g+1}) \arrow[dll] & \\
\mathbb{Z} \arrow[r, "\iota"'] & \mathbb{Z}^{2g} \oplus \mathbb{Z}^2 \arrow[r] & H_1(L_{g+1}) \arrow[r] & 0
\end{tikzcd}$$

By assumption, ι is the zero map. Hence $H_2(L_{g+1}) \simeq H_1(\partial L_{g,1}) \simeq \mathbb{Z}$ as ∂_2 is an isomorphism. Also, $\mathbb{Z}^{2g+2} \simeq H_1(L_{g+1})$ by exactness. □

7.9. Homology of non-orientable surfaces

Let M_1 be a triangulation of $\mathbb{RP}^2$. Let $M_{r,1}$ be M_r with a 2-simplex removed, so $\partial M_{r,1} \simeq \mathbb{S}^1$. Let $M_{r+1} = M_{r,1} \cup_{\partial M_{r,1}} M_{1,1}$. Then $M_{r+1,1} = M_{r,1} \cup_{\Delta^1} M_{1,1}$, attaching along an interval. For example, $|M_{1,1}|$ is homeomorphic to the Möbius band. Then $M_{r,1} \sim \bigvee_{i=1}^r S^1$.

Proposition.

$$H_k(M_r) = \begin{cases} \mathbb{Z}^{r-1} \oplus \mathbb{Z}/2\mathbb{Z} & k = 1 \\ \mathbb{Z} & k = 0 \\ 0 & \text{otherwise} \end{cases}$$

and

$$H_k(M_{r,1}) = \begin{cases} \mathbb{Z}^r & k = 1 \\ \mathbb{Z} & k = 0 \\ 0 & \text{otherwise} \end{cases}$$

Further, $\iota_{r\star} : H_1(\partial M_{r,1}) \to H_1(M_{r,1})$ has the property that $\iota_{r\star}(1)$ is twice a primitive element, or equivalently, $H_1(M_{r,1})/\operatorname{Im} \iota_{r\star} = \mathbb{Z}^{r-1} \oplus \mathbb{Z}/2\mathbb{Z}$.

Proof. We proceed by induction in the same way. For the base case, note that $\partial M_{1,1} \simeq S^1$ and $M_{1,1} \simeq S^1$, and the map from $\partial M_{1,1} \to M_{1,1}$ is given by $z \mapsto z^2$, so the map $H_1(S^1) \to H_1(S^1)$ is given by multiplication by 2. Suppose the result holds for $H_k(M_r)$. Then, $M_r = M_{r,1} \cup_{\partial M_{r,1}} \boldsymbol{\Delta}^2$, and

$$\begin{tikzcd}
 & H_2(M_{r,1}) \oplus H_2(\boldsymbol{\Delta}^2) \arrow[r] & H_2(M_r) \arrow[dll] & \\
H_1(\partial M_{r,1}) \arrow[r, "\iota_{r\star}"'] & H_1(M_{r,1}) \oplus H_1(\boldsymbol{\Delta}^2) \arrow[r] & H_1(M_r) \arrow[r] & 0
\end{tikzcd}$$

$\iota_{r\star}$ is injective, so $\partial_2 = 0$, giving $0 \longrightarrow H_2(M_r) \longrightarrow 0$. Hence,

$$\begin{tikzcd}
& 0 \oplus 0 \arrow[r] & 0 \arrow[dll, rounded corners, to path={ -- ([xshift=2ex]\tikztostart.east) |- ([yshift=2ex]\tikztotarget.north west) \tikztonodes -- (\tikztotarget.west)}] & \\
\mathbb{Z} \arrow[r, "\iota_{r\star}"'] & H_1(M_{r,1}) \oplus 0 \arrow[r] & \mathbb{Z}^{r-1} \oplus \mathbb{Z}/2\mathbb{Z} \arrow[r] & 0
\end{tikzcd}$$

Since $H_1(M_{r,1})$ is torsion-free,

$$\begin{tikzcd}
0 \arrow[r] & \mathbb{Z} \arrow[r] & H_1(M_{r,1}) \arrow[r] & \mathbb{Z}^{r-1} \oplus \mathbb{Z}/2\mathbb{Z} \arrow[r] & 0
\end{tikzcd}$$

gives that $H_1(M_{r,1}) = \mathbb{Z}^r$.

Now, $M_{r+1} = M_{r,1} \cup_{\partial M_{r,1}} M_{1,1}$ hence

$$\begin{tikzcd}
& H_2(M_{r,1}) \oplus H_2(M_{1,1}) \arrow[r] & H_2(M_{r+1}) \arrow[dll, rounded corners, to path={ -- ([xshift=2ex]\tikztostart.east) |- ([yshift=2ex]\tikztotarget.north west) \tikztonodes -- (\tikztotarget.west)}] & \\
H_1(\mathbb{S}^1) \arrow[r] & H_1(M_{r,1}) \oplus H_1(M_{r,1}) \arrow[r] & H_1(M_{r+1}) \arrow[r] & 0
\end{tikzcd}$$

so

$$\begin{tikzcd}
& 0 \oplus 0 \arrow[r] & 0 \arrow[dll, "0", rounded corners, to path={ -- ([xshift=2ex]\tikztostart.east) |- ([yshift=2ex]\tikztotarget.north west) \tikztonodes -- (\tikztotarget.west)}] & \\
\mathbb{Z} \arrow[r] & H_1(M_{r,1}) \oplus H_1(M_{r,1}) \arrow[r] & H_1(M_{r+1}) \arrow[r] & 0
\end{tikzcd}$$

Hence $H_1(M_{r+1}) \simeq \left. \mathbb{Z}^2 \oplus \mathbb{Z} \middle/ (2e_1, 2) \right. \simeq \mathbb{Z}^r \oplus \mathbb{Z}/2\mathbb{Z}$. □

7.10. Lefschetz fixed point theorem

Let (C, d) be a chain complex over $\mathbb{Q}$ (or any other field). Then $H_\bullet(C)$ is a $\mathbb{Q}$-vector space. Let $f : C \to C$ be a chain map, so it induces $f_\star : H_\bullet(C) \to H_\bullet(C)$. f and $f_\star$ are both linear endomorphisms of vector spaces.

Definition. The *Lefschetz number* of f is $L(f) = \sum_k (-1)^k \operatorname{tr} f_k$ where $f_k : C_k \to C_k$, and $L(f_\star) = \sum_k (-1)^k \operatorname{tr} f_{k\star}$ where $f_{k\star} : H_k(C) \to H_k(C)$.

Proposition. $L(f) = L(f_\star)$.

Proof. Let $U_k = \operatorname{Im} d_{k+1} \subseteq \ker d_k \subseteq C_k$. Then, $\ker d_k = U_k \oplus V_k$, and $C_k = U_k \oplus V_k \oplus U_k'$. Then $d : U_k' \to U_{k-1}$ is an isomorphism. With respect to this decomposition, d is a matrix in block form given by

$$d = \begin{pmatrix} 0 & 0 & I \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Also, $f(\operatorname{Im} d_{k+1}) \subseteq \operatorname{Im} d_{k+1}$ since f is a chain map, and $f(\ker d_k) \subseteq \ker d_k$. So in block form,

$$f = \begin{pmatrix} A_k & X_k & \star \\ 0 & B_k & \star \\ 0 & 0 & A'_k \end{pmatrix}$$

Then, the equation $df = fd$ shows $A_k = A'_{k+1}$. Hence, $H_k(C) = {}^{\ker d_k}\!\big/_{\operatorname{Im} d_{k+1}} = {}^{U_k \oplus V_k}\!\big/_{U_k} \simeq V_k$, and $f_{k\star} : H_k(C) \to H_k(C)$ maps $[v]$ to $[B_k v + X_k v] = [B_k v]$, so $f_{k\star}$ is multiplication by B_k. Then $L(f) = \sum_k (-1)^k \operatorname{tr} f_k = \sum (-1)^k (\operatorname{tr} A_k + \operatorname{tr} B_k + \operatorname{tr} A_{k-1}) = \sum (-1)^k \operatorname{tr} B_k = L(f_\star)$. □

Definition. Let $C = C_\bullet(K)$. Then the *Euler characteristic* is defined by $\chi(C) = L(\mathrm{id}_C)$. Hence $\chi(C(K)) = \sum_k (-1)^k \dim C_k(K)$. Note that $L(\mathrm{id}_C) = L(\mathrm{id}_{H_\bullet(K)}) = \sum_k (-1)^k \dim H_k(K)$ depends only on $|K|$.

Theorem (Lefschetz fixed point theorem). Let $F : |K| \to |K|$ be a continuous map. Let $L(F) = L(F_\star)$ be the Lefschetz number of F, where $F_\star : H_\bullet(K) \to H_\bullet(K)$. Then if $L(F) \neq 0$, F has a fixed point.

Remark. This is a generalisation of the Brouwer fixed point theorem.

Proof sketch. If F has no fixed point, then since $|K|$ is compact, there exists $\varepsilon > 0$ such that $|F(x) - x| \geq \varepsilon$ for all x. If $f : B^{r+n}K \to B^r K$ is a simplicial approximation of F, then the above implies that $F_\star(\sigma)$ does not contain σ for any simplex $\sigma \in C_\bullet(K)$. Hence $L(F) = L(f) = 0$. □

II. Probability and Measure

Lectured in Michaelmas 2022 by PROF. R. NICKL

In this course, we study measure theory and integration, and its applications to probability theory. We begin by defining the notion of a measure, which extends the notion of the length of an interval to a much larger class of 'measurable' sets. In the context of a probability space, a probability measure is a way to associate probabilities to events that could occur. Measures have the countable additivity property, which allows us to compute the measure of certain limits of measurable sets. Using this property, we can analyse limiting behaviour by considering the measure of a set on which a certain event occurs.

Measure theory allows us to define the Lebesgue integral. This integral agrees with the Riemann integral on most well-behaved functions, but it has many more convenient properties concerning limits. For example, the dominated convergence theorem gives a sufficient condition for when the limit of the integrals of functions is the integral of the limit. Another example is that the set of Lebegsue integrable functions forms a complete normed vector space, but this is not true of the Riemann integral.

Using the Lebesgue integral, we can define the Fourier transform of an integrable function. This linear operator is 'almost' injective: if the Fourier transform of a function is also integrable, we can recover the original function almost everywhere. Properties of the Fourier transform are used to deduce the central limit theorem.

Contents

1. Measures

1.1. Definitions

Definition. Let E be a (nonempty) set. A collection $\mathcal{E}$ of subsets of E is called a *σ-algebra* if the following properties hold:

- $\varnothing \in \mathcal{E}$;
- $A \in \mathcal{E} \implies A^c = E \setminus A \in \mathcal{E}$;
- if $(A_n)_{n\in\mathbb{N}}$ is a countable collection of sets in $\mathcal{E}$, $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{E}$.

Example. Let $\mathcal{E} = \{\varnothing, E\}$. This is a σ-algebra. Also, $\mathcal{P}(E) = \{A \subseteq E\}$ is a σ-algebra.

Remark. Since $\bigcap_n A_n = \left(\bigcup_n A_n^c\right)^c$, any σ-algebra $\mathcal{E}$ is closed under countable intersections as well as under countable unions. Note that $B \setminus A = B \cap A^c \in \mathcal{E}$, so σ-algebras are closed under set difference.

Definition. A set E with a σ-algebra $\mathcal{E}$ is called a *measurable space*. The elements of $\mathcal{E}$ are called *measurable sets*.

Definition. A *measure* μ is a set function $\mu : \mathcal{E} \to [0,\infty]$, such that $\mu(\varnothing) = 0$, and for a sequence $(A_n)_{n\in\mathbb{N}}$ such that the A_n are disjoint, we have

$$\mu\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mu(A_n)$$

This is the *countable additivity* property of the measure.

Remark. If E is countable, then for any $A \in \mathcal{P}(E)$ and measure μ, we have

$$\mu(A) = \mu\left(\bigcup_{x\in A} \{x\}\right) = \sum_{x\in A} \mu(\{x\})$$

Hence, measures are uniquely defined by the measure of each singleton. This corresponds to the notion of a probability mass function.

Definition. For a collection $\mathcal{A}$ of subsets of E, we define the σ-algebra $\sigma(\mathcal{A})$ *generated by* $\mathcal{A}$ by

$$\sigma(\mathcal{A}) = \{A \subseteq E : A \in \mathcal{E} \text{ for all } \sigma\text{-algebras } \mathcal{E} \supseteq \mathcal{A}\}$$

So it is the smallest σ-algebra containing $\mathcal{A}$. Equivalently,

$$\sigma(\mathcal{A}) = \bigcap_{\mathcal{E}\supseteq\mathcal{A}, \mathcal{E} \text{ a } \sigma\text{-algebra}} \mathcal{E}$$

1.2. Rings and algebras

To construct good generators, we define the following.

Definition. $\mathcal{A} \subseteq \mathcal{P}(E)$ is called a *ring* over E if $\varnothing \in \mathcal{A}$ and $A, B \in \mathcal{A}$ implies $B \setminus A \in \mathcal{A}$ and $A \cup B \in \mathcal{A}$.

Rings are easier to manage than σ-algebras because there are only finitary operators.

Definition. $\mathcal{A}$ is called an *algebra* over E if $\varnothing \in \mathcal{A}$ and $A, B \in \mathcal{A}$ implies $A^c \in \mathcal{A}$ and $A \cup B \in \mathcal{A}$.

Remark. Rings are closed under symmetric difference $A \triangle B = (B \setminus A) \cup (A \setminus B)$, and are closed under intersections $A \cap B = A \cup B \setminus A \triangle B$. Algebras are rings, because $B \setminus A = B \cap A^c = (B^c \cup A)^c$. Not all rings are algebras, because rings do not need to include the entire space.

Proposition (Disjointification of countable unions)**.** Consider $\bigcup_n A_n$ for $A_n \in \mathcal{E}$, where $\mathcal{E}$ is a σ-algebra (or a ring, if the union is finite). Then there exist $B_n \in \mathcal{E}$ that are disjoint such that $\bigcup_n A_n = \bigcup_n B_n$.

Proof. Define $\widetilde{A}_n = \bigcup_{j \leq n} A_j$, then $B_{n+1} = \widetilde{A}_n \setminus \widetilde{A}_{n-1}$. □

Definition. A *set function* on a collection $\mathcal{A}$ of subsets of E, where $\varnothing \in \mathcal{A}$, is a map $\mu : \mathcal{A} \to [0, \infty]$ such that $\mu(\varnothing) = 0$. We say μ is *increasing* if $\mu(A) \leq \mu(B)$ for all $A \subseteq B$ in $\mathcal{A}$. We say μ is *additive* if $\mu(A \cup B) = \mu(A) + \mu(B)$ for disjoint $A, B \in \mathcal{A}$ and $A \cup B \in \mathcal{A}$. We say μ is *countably additive* if $\mu\big(\bigcup_n A_n\big) = \sum_n \mu(A_n)$ for disjoint sequences A_n where $\bigcup_n A_n$ and each A_n lie in $\mathcal{A}$. We say μ is *countably subadditive* if $\mu\big(\bigcup_n A_n\big) \leq \sum_n \mu(A_n)$ for arbitrary sequences A_n under the above conditions.

Remark. A measure satisfies all four of the above conditions. Countable additivity implies the other conditions.

Theorem (Carathéodory's theorem)**.** Let μ be a countably additive set function on a ring $\mathcal{A}$ of subsets of E. Then there exists a measure $\mu^\star$ on $\sigma(\mathcal{A})$ such that $\mu^\star|_{\mathcal{A}} = \mu$.

We will later prove that this extended measure is unique.

Proof. For $B \subseteq E$, we define the *outer measure* $\mu^\star$ as

$$\mu^\star(B) = \inf\left\{ \sum_{n \in \mathbb{N}} \mu(A_n), A_n \in \mathcal{A}, B \subseteq \bigcup_{n \in \mathbb{N}} A_n \right\}$$

If there is no sequence A_n such that $B \subseteq \bigcup_{n \in \mathbb{N}} A_n$, we declare the outer measure $\mu^\star(B)$ to be ∞. We define the class

$$\mathcal{M} = \{A \subseteq E \mid \forall B \subseteq E, \mu^\star(B) = \mu^\star(B \cap A) + \mu^\star(B \cap A^c)\}$$

This is the class of *$\mu^\star$-measurable sets.*

Step 1. $\mu^\star$ is countably sub-additive on $\mathcal{P}(E)$. It suffices to prove that for $B \subseteq E$ and $B_n \subseteq E$ such that $B \subseteq \bigcup_n B_n$ we have

$$\mu^\star(B) \leq \sum_n \mu^\star(B_n) \tag{†}$$

We can assume without loss of generality that $\mu^\star(B_n) < \infty$ for all n, otherwise there is nothing to prove. For all $\varepsilon > 0$ there exists a collection $A_{n,m}$ such that $B_n \subseteq \bigcup_m A_{n,m}$ and

$$\mu^\star(B_n) + \frac{\varepsilon}{2^n} \geq \sum_m \mu(A_{n,m})$$

Now, since $\mu^\star$ is increasing, and $B \subseteq \bigcup_n B_n \subseteq \bigcup_n \bigcup_m A_{n,m}$, we have

$$\mu^\star(B) \leq \mu^\star\left(\bigcup_{n,m} A_{n,m}\right) \leq \sum_{n,m} \mu(A_{n,m}) \leq \sum_n \mu^\star(B_n) + \sum_n \frac{\varepsilon}{2^n} = \sum_n \mu^\star(B_n) + \varepsilon$$

Since ε was arbitrary in the construction, (†) follows by construction.

Step 2. $\mu^\star$ extends μ. Let $A \in \mathcal{A}$, and we want to show $\mu^\star(A) = \mu(A)$. We can write $A = A \cup \varnothing \cup \dots$, hence $\mu^\star(A) \leq \mu(A) + 0 + \dots = \mu(A)$ by definition of $\mu^\star$. We need to prove the converse, that $\mu(A) \leq \mu^\star(A)$. If $\mu^\star$ is infinite, there is nothing to prove. For the finite case, suppose there is a sequence A_n where $\mu(A_n) < \infty$ and $A \subseteq \bigcup_n A_n$. Then, $A = \bigcup_n (A \cap A_n)$, which is a union of elements of the ring $\mathcal{A}$. Since μ is a countably additive set function on $\mathcal{A}$, it is countably subadditive. Hence $\mu(A) \leq \sum_n \mu(A \cap A_n) \leq \sum_n \mu(A_n)$. Since the A_n were arbitrary, we have $\mu(A) \leq \mu^\star(A)$ as required.

Step 3. $\mathcal{M} \supseteq \mathcal{A}$. Let $A \in \mathcal{A}$. We must show that for all $B \subseteq E$, $\mu^\star(B) = \mu^\star(B \cap A) + \mu^\star(B \cap A^c)$. We have $B \subseteq (B \cap A) \cup (B \cap A^c) \cup \varnothing \cup \dots$, hence by countable subadditivity (†), $\mu^\star(B) \leq \mu^\star(B \cap A) + \mu^\star(B \cap A^c)$. It now suffices to prove the converse, that $\mu^\star(B) \geq \mu^\star(B \cap A) + \mu^\star(B \cap A^c)$. We can assume $\mu^\star(B)$ is finite, and assume there exists $A_n \in \mathcal{A}$ such that $B \subseteq \bigcup_n A_n$ and $\mu^\star(B) + \varepsilon \geq \sum_n \mu(A_n)$. Now, $B \cap A \subseteq \bigcup_n (A_n \cap A)$, and $B \cap A^c \subseteq \bigcup_n (A_n \cap A^c)$. All of the members of these two unions are elements of $\mathcal{A}$, since $A_n \cap A^c = A_n \setminus A$. Therefore,

$$\begin{aligned}
\mu^\star(B \cap A) + \mu^\star(B \cap A^c) &\leq \sum_n \mu(A_n \cap A) + \sum_n \mu(A_n \cap A^c) \\
&\leq \sum_n [\mu(A_n \cap A) + \mu(A_n \cap A^c)] \\
&\leq \sum_n \mu(A_n) \leq \mu^\star(B) + \varepsilon
\end{aligned}$$

Since ε was arbitrary, $\mu^\star(B) = \mu^\star(B \cap A) + \mu^\star(B \cap A^c)$ as required.

Step 4. $\mathcal{M}$ is an algebra. Clearly $\varnothing$ lies in $\mathcal{M}$, and by the symmetry in the definition of $\mathcal{M}$, complements lie in $\mathcal{M}$. We need to check $\mathcal{M}$ is stable under finite intersections. Let $A_1, A_2 \in \mathcal{M}$ and let $B \subseteq E$. We have

$$\mu^\star(B) = \mu^\star(B \cap A_1) + \mu^\star(B \cap A_1^c) = \mu^\star(B \cap A_1 \cap A_2) + \mu^\star(B \cap A_1 \cap A_2^c) + \mu^\star(B \cap A_1^c)$$

We can write $A_1 \cap A_2^c = (A_1 \cap A_2^c)^c \cap A_1$, and $A_1^c = (A_1 \cap A_2)^c \cap A_1^c$. Hence

$$\begin{aligned}\mu^\star(B) &= \mu^\star(B \cap A_1 \cap A_2) + \mu^\star(B \cap (A_1 \cap A_2)^c \cap A_1) + \mu^\star(B \cap (A_1 \cap A_2)^c \cap A_1^c) \\ &= \mu^\star(B \cap A_1 \cap A_2) + \mu^\star(B \cap (A_1 \cap A_2)^c)\end{aligned}$$

which is the requirement for $A_1 \cap A_2$ to lie in $\mathcal{M}$.

Step 5. $\mathcal{M}$ is a σ-algebra and $\mu^\star$ is a measure on $\mathcal{M}$. It suffices now to show that $\mathcal{M}$ has countable unions and the measure respects these countable unions. Let $A = \bigcup_n A_n$ for $A_n \in \mathcal{M}$. Without loss of generality, let the A_n be disjoint. We want to show $A \in \mathcal{M}$, and that $\mu^\star(A) = \sum_n \mu^\star(A_n)$. By (†), we have $\mu^\star(B) \leq \mu^\star(B \cap A) + \mu^\star(B \cap A^c) + 0 + \dots$ so we need to check only the converse of this inequality. Also, $\mu^\star(A) \leq \sum_n \mu^\star(A_n)$, so we need only check the converse of this inequality as well. Similarly to before,

$$\begin{aligned}\mu^\star(B) &= \mu^\star(B \cap A_1) + \mu^\star(B \cap A_1^c) \\ &= \mu^\star(B \cap A_1) + \mu^\star(B \cap A_1^c \cap A_2) + \mu^\star(B \cap A_1^c \cap A_2^c) \\ &= \mu^\star(B \cap A_1) + \mu^\star(B \cap A_2) + \mu^\star(B \cap A_1^c \cap A_2^c) \\ &= \mu^\star(B \cap A_1) + \mu^\star(B \cap A_2) + \mu^\star(B \cap A_1^c \cap A_2^c \cap A_3) + \mu^\star(B \cap A_1^c \cap A_2^c \cap A_3^c) \\ &= \mu^\star(B \cap A_1) + \mu^\star(B \cap A_2) + \mu^\star(B \cap A_3) + \mu^\star(B \cap A_1^c \cap A_2^c \cap A_3^c) \\ &= \cdots \\ &= \sum_{n \leq N} \mu^\star(B \cap A_n) + \mu^\star(B \cap A_1^c \cap \dots \cap A_N^c)\end{aligned}$$

Since $\bigcup_{n \leq N} A_n \subseteq A$, we have $\bigcap_{n \leq N} A_n^c \supseteq A^c$. $\mu^\star$ is increasing, hence, taking limits,

$$\mu^\star(B) \geq \sum_{n=1}^\infty \mu^\star(B \cap A_n) + \mu^\star(B \cap A^c)$$

By (†),

$$\mu^\star(B) \geq \mu^\star(B \cap A) + \mu^\star(B \cap A^c)$$

as required. Hence $\mathcal{M}$ is a σ-algebra. For the other inequality, we take the above result for $B = A$.

$$\mu^\star(A) \geq \sum_{n=1}^\infty \mu^\star(A \cap A_n) + \mu^\star(A \cap A^c) = \sum_{n=1}^\infty \mu^\star(A_n)$$

So $\mu^\star$ is countably additive on $\mathcal{M}$ and is hence a measure on $\mathcal{M}$. □

1.3. Uniqueness of extension

Definition. A collection $\mathcal{A}$ of subsets of E is called a *π-system* if $\varnothing \in \mathcal{A}$ and $A, B \in \mathcal{A} \implies A \cap B \in \mathcal{A}$.

Definition. A collection $\mathcal{A}$ of subsets of E is called a *d-system* if $E \in \mathcal{A}$, and if $B_1 \subset B_2$ are elements of $\mathcal{A}$, we have $B_2 \setminus B_1 \in \mathcal{A}$, and if $A_n \in \mathcal{A}$ and A_n is an increasing sequence of sets, we have $\bigcup_n A_n \in \mathcal{A}$.

Proposition. A d-system which is also a π-system is a σ-algebra.

Proof. Refer to the first example sheet. □

Lemma (Dynkin)**.** Let $\mathcal{A}$ be a π-system. Then any d-system that contains $\mathcal{A}$ also contains $\sigma(\mathcal{A})$.

Proof. We define

$$\mathcal{D} = \bigcap_{\mathcal{D}' \text{ is a } d\text{-system; } \mathcal{D}' \supseteq \mathcal{A}} \mathcal{D}'$$

We can show this is a d-system. It suffices to prove that $\mathcal{D}$ is a π-system, because this is then a σ-algebra. We now define

$$\mathcal{D}' = \{B \in \mathcal{D} \mid \forall A \in \mathcal{A}, B \cap A \in \mathcal{D}\}$$

We can see that $\mathcal{D}' \supseteq \mathcal{A}$, as $\mathcal{A}$ is a π-system. We now show that $\mathcal{D}'$ is a d-system. Clearly $E \cap A = A \in \mathcal{A} \subseteq \mathcal{D}'$ hence $E \in \mathcal{D}'$. Let $B_1, B_2 \in \mathcal{D}'$ such that $B_1 \subseteq B_2$. Then $(B_2 \setminus B_1) \cap A = (B_2 \cap A) \setminus (B_1 \cap A)$, and since $B_i \cap A \in \mathcal{D}$ this difference also lies in $\mathcal{D}$, so $B_2 \setminus B_1 \in \mathcal{D}'$. Now, suppose B_n is an increasing sequence converging to B, and $B_n \in \mathcal{D}'$. Then $B_n \cap A \in \mathcal{D}$, and $\mathcal{D}$ is a d-system, we have $B \cap A \in \mathcal{D}$, so $B \in \mathcal{D}'$.

Hence $\mathcal{D}'$ is a d-system that contains $\mathcal{A}$, so $\mathcal{D} \subseteq \mathcal{D}'$, and $\mathcal{D}' \subseteq \mathcal{D}$ by construction of $\mathcal{D}'$, giving $\mathcal{D} = \mathcal{D}'$. We then define

$$\mathcal{D}'' = \{B \in \mathcal{D} \mid \forall A \in \mathcal{D}, B \cap A \in \mathcal{D}\}$$

Note that $\mathcal{A} \subseteq \mathcal{D}''$, because $\mathcal{D}' = \mathcal{D} \supseteq \mathcal{A}$. Running the same argument as before, we can show that $\mathcal{D}'' = \mathcal{D}$, and so $\mathcal{D}'' = \mathcal{D}$ is a π-system. □

Theorem (Uniqueness of extension)**.** Let μ_1, μ_2 be measures on a measurable space $(E, \mathcal{E})$, such that $\mu_1(E) = \mu_2(E) < \infty$. Suppose that μ_1 and μ_2 coincide on a π-system $\mathcal{A}$, such that $\mathcal{E} \subseteq \sigma(\mathcal{A})$. Then $\mu_1 = \mu_2$ on $\sigma(\mathcal{A})$, and hence on $\mathcal{E}$.

Proof. We define

$$\mathcal{D} = \{A \in \mathcal{E} \mid \mu_1(A) = \mu_2(A)\}$$

This collection contains $\mathcal{A}$ by assumption. By Dynkin's lemma, it suffices to prove $\mathcal{D}$ is a d-system, because then $\mathcal{D} \supseteq \sigma(\mathcal{A}) \supseteq \mathcal{E}$ giving $\mathcal{D} = \mathcal{E}$. Note that $E \in \mathcal{D}$ by assumption. By additivity and finiteness of μ_i, for $B_1 \subseteq B_2$ elements of $\mathcal{D}$, we have $\mu_1(B_2 \setminus B_1) = \mu_1(B_2) - \mu_1(B_1) = \mu_2(B_2) - \mu_2(B_1) = \mu_2(B_2 \setminus B_1)$, where the subtractions are valid by finiteness of μ, so set differences lie in $\mathcal{D}$.

Now suppose B_n is an increasing sequence converging to B for $B_n \in \mathcal{D}$. This implies that $B \setminus B_n$ is a decreasing sequence converging to $\varnothing$, and by a result from the first example sheet we have $\mu_i(B \setminus B_n) \to \mu(\varnothing) = 0$. Since μ_i are finite, $\mu_i(B_n) \to \mu_i(B)$ as $n \to \infty$. Then, $\mu_1(B) = \lim_{n \in \mathbb{N}} \mu_1(B_n) = \lim_{n \in \mathbb{N}} \mu_2(B_n) = \mu_2(B)$, so $\mathcal{D}$ is closed under increasing sequences and hence is a d system. □

Remark. The above theorem applies to finite measures (μ such that $\mu(E) < \infty$) only. However, the theorem can be extended to measures that are σ-finite, for which $E = \bigcup_{n\in\mathbb{N}} E_n$ where $\mu(E_n) < \infty$.

1.4. Borel measures

Definition. Let (E, τ) be a Hausdorff topological space. The σ-algebra generated by the open sets of E is called the *Borel σ-algebra* on E, denoted $\mathcal{B}(E) = \sigma(\tau)$. We write $\mathcal{B} = \mathcal{B}(\mathbb{R})$. Members of $\mathcal{B}(E)$ are called *Borel sets*. A measure μ on $(E, \mathcal{B}(E))$ is called a *Borel measure on E*. A *Radon measure* is a Borel measure μ on E such that $\mu(K) < \infty$ for all $K \subseteq E$ compact. Note that in a Hausdorff space, compact sets are closed and hence measurable.

1.5. Lebesgue measure

We will construct a unique Borel measure μ on $\mathbb{R}^d$ such that

$$\mu\left(\prod_{i=1}^{d}[a_i, b_i]\right) = \prod_{i=1}^{d} |b_i - a_i|$$

Initially, we will perform this construction for $d = 1$, and later we will consider product measures to extend this to higher dimensions.

Theorem (Construction of the Lebesgue measure)**.** There exists a unique Borel measure μ on $\mathbb{R}$ such that

$$a < b \implies \mu((a, b]) = b - a$$

Proof. Consider the subsets of $\mathbb{R}$ of the form

$$A = (a_1, b_1] \cup \cdots \cup (a_n, b_n]$$

where the intervals in question are disjoint. The set $\mathcal{A}$ of such sets forms a ring and a π-system of Borel sets. This generates the same σ-algebra as that generated by finite unions of open intervals, by the first example sheet. Open intervals with rational endpoints generate $\mathcal{B}$, so $\sigma(A) \supseteq \mathcal{B}$. We define the set function μ on $\mathcal{A}$ by $\mu(A) = \sum_{i=1}^{n}(b_i - a_i)$. μ is additive, and well-defined since if $A = \bigcup_j C_j = \bigcup_k D_k$ for distinct disjoint unions, we can write $C_j = \bigcup_k (C_j \cap D_k)$ and $D_k = \bigcup_j (D_k \cap C_j)$, giving

$$\mu(A) = \mu\left(\bigcup_j C_j\right) = \sum_j \mu(C_j) = \sum_j \mu\left(\bigcup_j (C_j \cap D_k)\right) = \sum_j \sum_k \mu(C_j \cap D_k) = \mu\left(\bigcup_k D_k\right)$$

To prove the existence of μ on $\mathcal{B}$, we apply Carathéodory's extension theorem, and therefore must check that μ is countably additive on $\mathcal{A}$. Equivalently, by a question on an example sheet, it suffices to show that for all sequences $A_n \in \mathcal{A}$ such that A_n decreases to $\varnothing$, we have $\mu(A_n) \to 0$. Suppose this is not the case, so there exist $\varepsilon > 0$ and $B_n \in \mathcal{A}$ such that

B_n decreases to $\varnothing$ but $\mu(B_n) \geq 2\varepsilon$ for infinitely many n (and so without loss of generality for all n). We can approximate B_n from within by a sequence C_n. Suppose $B_n = \bigcup_{i=1}^{N_n}(a_{ni}, b_{ni}]$, then define $C_n = \bigcup_{i=1}^{N_n}(a_{ni} + \frac{2^{-n}\varepsilon}{N_n}, b_{ni}]$. Note that the C_n lie in $\mathcal{A}$, and $\mu(B_n \setminus C_n) \leq 2^{-n}\varepsilon$ Since B_n is decreasing, we have $B_N = \bigcap_{n\leq N} B_n$, and

$$B_N \setminus (C_1 \cap \dots \cap C_N) = B_n \cap \left(\bigcup_{n\leq N} C_n^c \right) = \bigcup_{n\leq N} B_N \setminus C_n \subseteq \bigcup_{n\leq N} B_n \setminus C_n$$

Since μ is increasing,

$$\mu(B_N \setminus (C_1 \cap \dots \cap C_N)) \leq \mu\left(\bigcup_{n\leq N} B_n \setminus C_n \right) \leq \sum_{n\leq N} \mu(B_n \setminus C_n) \leq \sum_{n\leq N} 2^{-N}\varepsilon \leq \varepsilon$$

Since in addition $\mu(B_N) \geq 2\varepsilon$, additivity implies that $\mu(C_1 \cap \dots \cap C_N) \geq \varepsilon$. This means that $C_1 \cap \dots \cap C_N$ cannot be empty. We can add the left endpoints of the intervals, giving $K_N = \overline{C}_1 \cap \dots \cap \overline{C}_N$. By Analysis I, K_N is a nested sequence of nonempty closed intervals and therefore there is a point $x \in \mathbb{R}$ such that $x \in K_N$ for all N. But $K_N \subseteq \overline{C}_N \subseteq B_N$, so $x \in \bigcap_N B_n$, which is a contradiction since $\bigcap_N B_N$ is empty. Therefore, a measure μ on $\mathcal{B}$ exists.

Now we prove uniqueness. Suppose μ, λ are measures such that the measure of an interval $(a, b]$ is $b - a$. We define new measures $\mu_n(A) = \mu(A \cap (n, n+1])$ and $\lambda_n(A) = \lambda(A \cap (n, n+1])$. These new measures are finite with total mass 1. Hence, we can use the uniqueness of extension theorem to show $\mu_n = \lambda_n$ on $\mathcal{B}$. We find

$$\mu(A) = \mu\left(\bigcup_n A \cap (n, n+1] \right) = \sum_{n\in\mathbb{Z}} \mu(A \cap (n, n+1]) = \sum_{n\in\mathbb{Z}} \mu_n(A) = \sum_{n\in\mathbb{Z}} \lambda_n(A) = \dots = \lambda(A)$$

□

Definition. A Borel set $B \in \mathcal{B}$ is called a *Lebesgue null set* if $\mu(B) = 0$.

Remark. A singleton $\{x\}$ can be written as $\bigcap_n \left(x - \frac{1}{n}, x\right]$, hence $\mu(x) = \lim_n \frac{1}{n} = 0$. Hence singletons are null sets. In particular, $\mu((a, b)) = \mu((a, b]) = \mu([a, b)) = \mu([a, b])$. Any countable set $Q = \bigcup_q \{q\}$ is a null set. Not all null sets are countable; the Cantor set is an example.

The Lebesgue measure is *translation-invariant*. Let $x \in \mathbb{R}$, then the set $B+x = \{b + x \mid b \in B\}$ lies in $\mathcal{B}$ if and only if $B \in \mathcal{B}$, and in this case, it satisfies $\mu(B + x) = \mu(B)$. We can define the translated Lebesgue measure $\mu_x(B) = \mu(B + x)$ for all $B \in \mathcal{B}$, but since the Lebesgue measure is unique, $\mu_x = \mu$.

The class of outer measurable sets $\mathcal{M}$ used in Carathéodory's extension theorem is here called the class of Lebesgue measurable sets. This class can be shown to be

$$\mathcal{M} = \{M = A \cup N, A \in \mathcal{B}, N \subseteq B, B \in \mathcal{B}, \mu(B) = 0\} \supsetneq \mathcal{B}$$

1.6. Existence of non-measurable sets

Assuming the axiom of choice, there exists a non-measurable set of reals. Consider $E = (0, 1]$ with addition defined modulo one. By the same argument as before, the Lebesgue measure is translation-invariant modulo one. Consider the subgroup $Q = E \cap \mathbb{Q}$ of $(E, +)$. We define $x \sim y$ if $x - y \in Q$. Then, this gives equivalence classes $[x] = \{y \in E : x \sim y\}$ for all $x \in E$. Assuming the axiom of choice, we can select a representative of $[x]$ for each $x \in E$, and denote by S the set of such representatives. We can partition E into the union of its cosets, so $E = \bigcup_{q \in Q}(S + q)$ is a disjoint union.

Suppose S is a Borel set. Then $S + q$ is also a Borel set. We can therefore write

$$1 = \mu(E) = \mu\left(\bigcup_{q \in Q}(S + q)\right) = \sum_{q \in Q} \mu(S + q) = \sum_{q \in Q} \mu(S)$$

But no value for $\mu(S) \in [0, \infty]$ can be assigned to make this equation hold. Therefore S is not a Borel set.

One can further show that μ cannot be extended to all subsets $\mathcal{P}(E)$.

Theorem (Banach, Kuratowski)**.** Assuming the continuum hypothesis, there exists no measure μ on the set $\mathcal{P}((0, 1])$ such that $\mu((0, 1]) = 1$ and $\mu(\{x\}) = 0$ for $x \in (0, 1]$.

1.7. Probability spaces

Definition. If a measure space $(E, \mathcal{E}, \mu)$ has $\mu(E) = 1$, we call it a *probability space*, and instead write $(\Omega, \mathcal{F}, \mathbb{P})$. We call Ω the outcome space or sample space, $\mathcal{F}$ the set of events, and $\mathbb{P}$ the probability measure.

The axioms of probability theory (Kolmogorov, 1933), are

(i) $\mathbb{P}(\Omega) = 1$;

(ii) $0 \leq \mathbb{P}(E) \leq 1$ for all $E \in \mathcal{F}$;

(iii) if A_n are a disjoint sequence of events in $\mathcal{F}$, then $\mathbb{P}\left(\bigcup_n A_n\right) = \sum_n \mathbb{P}(A_n)$.

This is exactly what is required by our definition: $\mathbb{P}$ is a measure on a σ-algebra.

Definition. Events $A_i, i \in I$ are *independent* if for all finite $J \subseteq I$, we have

$$\mathbb{P}\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} \mathbb{P}(A_j)$$

σ-algebras $\mathcal{A}_i, i \in I$ are independent if for any $A_j \in \mathcal{A}_j$ where $J \subseteq I$ is finite, the A_j are independent.

Kolmogorov showed that these definitions are sufficient to derive the law of large numbers.

Proposition. Let $\mathcal{A}_1, \mathcal{A}_2$ be π-systems of sets in $\mathcal{F}$. Suppose $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\,\mathbb{P}(A_2)$ for all $A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2$. Then the σ-algebras $\sigma(\mathcal{A}_1), \sigma(\mathcal{A}_2)$ are independent.

This follows by uniqueness.

1.8. Borel–Cantelli lemmas

Definition. Let $A_n \in \mathcal{F}$ be a sequence of events. Then the *limit superior* of A_n is

$$\limsup_n A_n = \bigcap_n \bigcup_{m \geq n} A_m = \{A_n \text{ infinitely often}\}$$

The *limit inferior* of A_n is

$$\liminf_n A_n = \bigcup_n \bigcap_{m \geq n} A_m = \{A_n \text{ eventually}\}$$

Lemma (First Borel–Cantelli lemma). Let $A_n \in \mathcal{F}$ be a sequence of events such that $\sum_n \mathbb{P}(A_n) < \infty$. Then $\mathbb{P}(A_n \text{ infinitely often}) = 0$.

Proof. For all n, we have

$$\mathbb{P}\left(\limsup_n A_n\right) = \mathbb{P}\left(\bigcap_n \bigcup_{m \geq n} A_m\right) \leq \mathbb{P}\left(\bigcup_{m \geq n} A_m\right) \leq \sum_{m \geq n} \mathbb{P}(A_m) \to 0$$

□

This proof did not require that $\mathbb{P}$ be a probability measure, just that it is a measure. Therefore, we can use this for arbitrary measures.

Lemma (Second Borel–Cantelli lemma). Let $A_n \in \mathcal{F}$ be a sequence of independent events, and $\sum_n \mathbb{P}(A_n) = \infty$. Then $\mathbb{P}(A_n \text{ infinitely often}) = 1$.

Proof. By independence, for all $N \geq n \in \mathbb{N}$ and using $1 - a \leq e^{-a}$, we find

$$\mathbb{P}\left(\bigcap_{m=n}^{N} A_m^c\right) = \prod_{m=n}^{N} (1 - \mathbb{P}(A_m)) \leq \prod_{m=n}^{N} e^{-\mathbb{P}(A_m)} = e^{-\sum_{m=n}^{N} \mathbb{P}(A_m)}$$

As $N \to \infty$, this approaches zero. Since $\bigcap_{m=n}^{N} A_m^c$ decreases to $\bigcap_{m=n}^{\infty} A_m^c$, by countable additivity we must have $\mathbb{P}\left(\bigcap_{m=n}^{\infty} A_m^c\right) = 0$. But then

$$\mathbb{P}(A_n \text{ infinitely often}) = \mathbb{P}\left(\bigcap_n \bigcup_{m \geq n} A_m\right) = 1 - \mathbb{P}\left(\bigcup_n \bigcap_{m \geq n} A_m^c\right) \geq 1 - \sum_n \mathbb{P}\left(\bigcap_{m \geq n} A_m^c\right) = 1$$

Hence this probability is equal to one. □

2. Measurable functions

2.1. Definition

Definition. Let $(E, \mathcal{E}), (G, \mathcal{G})$ be measurable spaces. A function $f : E \to G$ is called $\mathcal{E}$-$\mathcal{G}$-*measurable* if when $A \in \mathcal{G}$, we have $f^{-1}(A) \in \mathcal{E}$.

Informally, the preimage of a measurable set under a measurable function is measurable.

If $G = \mathbb{R}$ and $\mathcal{G} = \mathcal{B}$, we can just say that $f : (E, \mathcal{E}) \to G$ is measurable. Moreover, if E is a topological space and $\mathcal{E} = \mathcal{B}(E)$, we say f is Borel measurable.

Note that preimages f^{-1} commute with many set operations such as intersection, union, and complement. This implies that $\left\{f^{-1}(A) \mid A \in \mathcal{G}\right\}$ is a σ-algebra over E, and likewise, $\left\{A \mid f^{-1}(A) \in \mathcal{E}\right\}$ is a σ-algebra over G. Hence, if $\mathcal{A}$ is a collection of subsets of G generating $\mathcal{G}$ such that $f^{-1}(A) \in \mathcal{E}$ for all $A \in \mathcal{A}$, the class $\left\{A \mid f^{-1} \in \mathcal{E}\right\}$ is a σ-algebra that contains $\mathcal{A}$ and hence that contains $\mathcal{G}$. In particular, it suffices to check $f^{-1}(A) \in \mathcal{E}$ for all elements of a generator to conclude that f is measurable.

If $f : (E, \mathcal{E}) \to \mathbb{R}$, the collection $\mathcal{A} = \{(-\infty, y] : y \in \mathbb{R}\}$ generates $\mathcal{B}$ as is shown on the first example sheet. Hence f is measurable whenever $f^{-1}((-\infty, y]) = \{x \in E \mid f(x) \leq y\} \in \mathcal{E}$ for all $y \in \mathbb{R}$.

If E is a topological space and $\mathcal{E} = \mathcal{B}(E)$, then if $f : E \to \mathbb{R}$ is continuous, the preimages of open sets B are open, and hence Borel sets. The open sets in $\mathbb{R}$ generate the σ-algebra $\mathcal{B}$. Hence, continuous functions to the real line are measurable.

Example. Consider the indicator function $\mathbb{1}_A$ of a set A. This is measurable if and only if A is measurable, or equivalently $A \in \mathcal{E}$.

Example. The composition of measurable functions is measurable. Measurability is preserved under addition, multiplication, countable infimum, countable supremum, countable limit inferior, countable limit superior, and some other operations. Note that given a collection of maps $\{f_i : E \to (G, \mathcal{G}) \mid i \in I\}$, we can make them all measurable by taking $\mathcal{E}$ to be a large enough σ-algebra, for instance $\sigma\left(\left\{f_i^{-1}(A) \mid A \in \mathcal{G}, i \in I\right\}\right)$.

2.2. Monotone class theorem

Theorem. Let $\mathcal{A}$ be a π-system that generates the σ-algebra $\mathcal{E}$ over E. Let $\mathcal{V}$ be a vector space of bounded maps from E to $\mathbb{R}$ such that

(i) $\mathbb{1}_E \in \mathcal{V}$;

(ii) $\mathbb{1}_A \in \mathcal{V}$ for all $A \in \mathcal{A}$;

(iii) if f is bounded and $f_n \in \mathcal{V}$ are nonnegative functions that form an increasing sequence that converge pointwise to f on E, then $f \in \mathcal{V}$.

Then $\mathcal{V}$ contains all bounded measurable functions $f : E \to \mathbb{R}$.

Proof. Define $\mathcal{D} = \{A \in \mathcal{E} \mid \mathbb{1}_A \in \mathcal{V}\}$. This contains $\mathcal{A}$ by hypothesis, as well as E itself. We show $\mathcal{D}$ is a d-system, so that by Dynkin's lemma, $\mathcal{E} = \mathcal{D}$. Indeed, $E \in \mathcal{D}$ by assumption. For $A \subseteq B$ and $A, B \in \mathcal{D}$, we have $\mathbb{1}_{B \setminus A} = \mathbb{1}_B - \mathbb{1}_A$ which is well-defined and lies in $\mathcal{V}$ as $\mathcal{V}$ is a vector space. Finally, if $A_n \in \mathcal{D}$ increases to A, we have $\mathbb{1}_{A_n}$ increases pointwise to $\mathbb{1}_A$, which lies in $\mathcal{V}$ by the second hypothesis. Hence $\mathcal{E} = \mathcal{D}$.

Let $f : E \to \mathbb{R}$ be a bounded measurable function, which we will assume at first is nonnegative. We define

$$f_n = \sum_{j=0}^{n2^n} \frac{j}{2^n} \mathbb{1}_{A_{n_j}}; \quad A_{n_j} = \begin{cases} \left\{x \in E \mid \frac{j}{2^n} < f(x) \leq \frac{j+1}{2^n}\right\} = f^{-1}\left(\left(\frac{j}{2^n}, \frac{j+1}{2^n}\right]\right) \in \mathcal{E} & \text{if } j \neq n2^n \\ \{x \in E \mid n < f(x)\} = f^{-1}((n, \infty)) & \text{if } j = n2^n \end{cases}$$

Since f is bounded, for $n > \|f\|_\infty$, we have $f_n \leq f \leq f_n + 2^{-n}$. Hence $|f_n - f| \leq 2^{-n} \to 0$. By assumption, the limit of the f_n, which is exactly f, also lies in $\mathcal{V}$.

Now, by separating any bounded measurable function f into its positive and negative parts, we find that these two parts lie in $\mathcal{V}$, and so $f \in \mathcal{V}$ as required. □

2.3. Image measures

Definition. Let $f : (E, \mathcal{E}) \to (G, \mathcal{G})$ be a measurable function, and μ is a measure on $(E, \mathcal{E})$. Then the *image measure* $\nu = \mu \circ f^{-1}$ is obtained from assigning $\nu(A) = \mu(f^{-1}(A))$ for all $A \in \mathcal{G}$.

Lemma. Let $g : \mathbb{R} \to \mathbb{R}$ be an increasing, right-continuous function, and set $g(\pm\infty) = \lim_{z \to \pm\infty} g(z)$. On $I = (g(-\infty), g(+\infty))$ we define the *generalised inverse*

$$f(x) = \inf\{y \in \mathbb{R} \mid x \leq g(y)\}$$

for $x \in I$. Then f is increasing, left-continuous, and $f(x) \leq y$ if and only if $x \leq g(y)$ for all $x \in I, y \in \mathbb{R}$.

Remark. f and g form a Galois connection.

Proof. Let $J_x = \{y \in \mathbb{R} \mid x \leq g(y)\}$. Since $x > g(-\infty)$, J_x is nonempty and bounded below. Hence $f(x)$ is a well-defined real number. If $y \in J_x$, then $y' \geq y$ implies $y' \in J_x$ since g is increasing. Further, if y_n converges from the right to y, and all $y_n \in J_x$, we can take limits in $x \leq g(y_n)$ to find $x \leq \lim_n g(y_n) = g(y)$ since g is right-continuous. Hence $y \in J_x$. So $J_x = [f(x), \infty)$. Hence $f(x) \leq y \iff x \leq g(y)$ as required.

If $x \leq x'$, we have $J_x \supseteq J_{x'}$ by definition, so $f(x) \leq f(x')$. Similarly, if x_n converges from the left to x, we have $J_x = \bigcap_n J_{x_n}$, so $f(x_n) \to f(x)$ as $x_n \to x$. □

Theorem. Let $g : \mathbb{R} \to \mathbb{R}$ be an increasing, right-continuous function, and set $g(\pm\infty) = \lim_{z \to \pm\infty} g(z)$. Then there exists a unique Radon measure μ_g on $\mathbb{R}$ such that $\mu_g((a, b]) = g(b) - g(a)$ for all $a < b$. Further, all Radon measures can be obtained in this way.

Proof. We will show that the generalised inverse f as defined above is measurable. For all $z \in \mathbb{R}$, we find $f^{-1}((-\infty, z]) = \{x : f(x) \leq z\} = \{x : x \leq g(z)\} = [-g(\infty), g(z)]$ which is measurable. Since $\mathcal{B}$ is generated by these such sets, f is $\mathcal{B}(I)$-$\mathcal{B}$ measurable as required. Therefore, the image measure $\mu_g = \mu \circ f^{-1}$, where μ is the Lebesgue measure on I, exists. Then for any $-\infty < a < b < \infty$, we have

$$\begin{aligned}\mu_g((a,b]) &= \mu(f^{-1}((a,b])) \\ &= \mu(\{x : a < f(x) \leq f(b)\}) \\ &= \mu(\{x : g(a) < x \leq g(b)\}) \\ &= g(b) - g(a)\end{aligned}$$

This uniquely determines μ_g by the same argument as shown previously for the Lebesgue measure μ on $\mathbb{R}$. Since g maps into $\mathbb{R}$, $g(b) - g(a) \in \mathbb{R}$ so any compact set has finite measure as it is a subset of a closed bounded interval.

Conversely, let ν be a Radon measure on $\mathbb{R}$. Define

$$g(y) = \begin{cases} \nu((0,y]) & \text{if } y \geq 0 \\ -\nu((y,0]) & \text{if } y < 0 \end{cases}$$

This is an increasing function in y, since ν is a measure. Since we are using right-closed intervals, g is right-continuous. Finally, $\nu((a,b]) = g(b) - g(a)$ which can be seen by case analysis and additivity of the measure ν. By uniqueness as before, this characterises ν in its entirety. □

Remark. Such image measures μ_g are called *Lebesgue–Stieltjes measures*, where g is the *Stieltjes distribution*.

Example. The *Dirac measure at x*, written δ_x, is defined by

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

This has Stieltjes distribution $g(x) = \mathbb{1}_{[x,\infty)}$.

2.4. Random variables

Definition. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $(E, \mathcal{E})$ be a measurable space. An *E-valued random variable X* is an $\mathcal{F}$-$\mathcal{E}$ measurable map $X : \Omega \to E$. When $E = \mathbb{R}$ or $\mathbb{R}^d$ with the Borel σ-algebra, we simply call X a random variable or random vector.

The *law* or *distribution* μ_X of a random variable X is given by the image measure $\mu_X = \mathbb{P} \circ X^{-1}$. When E is the real line, this measure has a distribution function

$$F_X(z) = \mu_X((-\infty, z]) = \mathbb{P}(X^{-1}(-\infty, z]) = \mathbb{P}\left(\{\omega \in \Omega \mid X(\omega) \leq z\}\right) = \mathbb{P}\left(X \leq z\right)$$

This uniquely determines μ_X by the π-system argument given above.

Using the properties of measures, we can show that any distribution function satisfies:

(i) F_X is increasing;

(ii) F_X is right-continuous;

(iii) $\lim_{z\to-\infty} F_X(z) = \mu_X(\varnothing) = 0$;

(iv) $\lim_{z\to\infty} F_X(z) = \mu_X(\mathbb{R}) = \mathbb{P}(\Omega) = 1$.

Given any function F_X satisfying each property, we can obtain a random variable X on $(\Omega, \mathcal{F}, \mathbb{P}) = ((0,1), \mathcal{B}((0,1)), \mu)$ by $X(\omega) = \inf\{x \mid \omega \leq f(x)\}$, and then F_X is the distribution function of X.

Definition. Consider a countable collection $(X_i : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, \mathcal{E}))$ for $i \in I$. This collection of random variables is called *independent* if the σ-algebras $\sigma\big(\{X_i^{-1}(A) : A \in \mathcal{E}\}\big)$ are independent.

For $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B})$ we show on an example sheet that this is equivalent to the condition

$$\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \mathbb{P}(X_1 \leq x_1) \dots \mathbb{P}(X_n \leq x_n)$$

for all finite subsets $\{X_1, \dots, X_n\}$ of the X_i.

2.5. Constructing independent random variables

We now construct an infinite sequence of independent random variables with prescribed distribution functions on $(\Omega, \mathcal{F}, \mathbb{P}) = ((0,1), \mathcal{B}, \mu)$ with μ the Lebesgue measure on $(0,1)$. We start with Bernoulli random variables.

Any $\omega \in (0,1)$ has a binary representation given by $(\omega_i) \in \{0,1\}^{\mathbb{N}}$, which is unique if we exclude infinitely long tails of zeroes from the binary representation. We can then define the *nth Rademacher function* $R_n(\omega) = \omega_n$ which extracts the nth bit from the binary expansion. Since each R_n can be given as the sum of 2^{n-1} indicator functions on measurable sets, they are measurable functions and are hence random variables. Their distribution is given by $\mathbb{P}(R_n = 1) = \frac{1}{2} = \mathbb{P}(R_n = 0)$, so we have constructed Bernoulli random variables with parameter $\frac{1}{2}$. We show they are independent. For a finite set $(x_i)_{i=1}^n$,

$$\mathbb{P}(R_1 = x_1, \dots, R_n = x_n) = 2^{-n} = \mathbb{P}(R_1 = x_1) \dots \mathbb{P}(R_n = x_n)$$

Therefore, the R_n are all independent, so countable sequences of independent random variables indeed exist. Now, take a bijection $m : \mathbb{N}^2 \to \mathbb{N}$ and define $Y_{nk} = R_{m(n,k)}$, which are independent random variables. We can now define $Y_n = \sum_k 2^{-k} Y_{nk}$. This converges for all $\omega \in \Omega$ since $|Y_{nk}| \leq 1$, and these are still independent. We show the Y_n are uniform random variables, by showing the distribution coincides with the uniform distribution on the π-system of intervals $\left(\frac{i}{2^m}, \frac{i+1}{2^{m+1}}\right]$ for $i = 0, \dots, 2^m - 1$, which generates $\mathcal{B}$.

$$\mathbb{P}\left(Y_n \in \left(\frac{i}{2^m}, \frac{i+1}{2^m}\right]\right) = \mathbb{P}\left(\frac{i}{2^m} < \sum_k 2^{-k} Y_{nk} \leq \frac{i+1}{2^n}\right) = 2^{-m} = \mu\left(\frac{i}{2^m}, \frac{i+1}{2^{m+1}}\right]$$

Hence $\mu_{Y_n} = \mu|_{(0,1)}$ by the uniqueness theorem, and so we have constructed an infinite sequence of independent uniform random variables Y_n. If F_n are probability distribution functions, taking the generalised inverse, we see that the $F_n^{-1}(Y_n)$ are independent and have distribution function F_n.

2.6. Convergence of measurable functions

Definition. We say that a property defining a set $A \in \mathcal{E}$ holds *μ-almost everywhere* if $\mu(A^c) = 0$ for a measure μ on $\mathcal{E}$. If $\mu = \mathbb{P}$, we say a property holds *$\mathbb{P}$-almost surely* or *with probability one*, if $\mathbb{P}(A) = 1$.

Definition. If f_n and f are measurable functions on $(E, \mathcal{E}, \mu)$, we say f_n *converges to f μ-almost everywhere* if $\mu(\{x \in E \mid f_n(x) \nrightarrow f(x)\}) = 0$. We say f_n *converges to f in μ-measure* if for all $\varepsilon > 0$, $\mu(\{x \in E \mid |f_n(x) - f(x)| > \varepsilon\}) \to 0$ as $n \to \infty$. For random variables, we say $X_n \to X$ *$\mathbb{P}$-almost surely* or *in $\mathbb{P}$-probability*, written $X_n \to^p X$, respectively. If X_n, X take values in $\mathbb{R}$, we say $X_n \to X$ *in distribution*, written $X_n \to^d X$ if $\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$ at all points x for which the limit $x \mapsto \mathbb{P}(X \leq x)$ is continuous.

We can show that $X_n \to^p X \implies X^n \to^d X$.

Theorem. Let $f_n : (E, \mathcal{E}, \mu) \to \mathbb{R}$ be measurable functions. Then,

(i) if $\mu(E) < \infty$, then $f_n \to 0$ almost everywhere implies that $f_n \to 0$ in measure;

(ii) if $f_n \to 0$ in measure, $f_{n_k} \to 0$ almost everywhere on some subsequence.

Proof. Let $\varepsilon > 0$.

$$\mu(|f_n| < \varepsilon) \geq \mu\left(\bigcap_{m \geq n} \{|f_m| \leq \varepsilon\}\right)$$

The sequence $\left(\bigcap_{m \geq n} \{|f_m| \leq \varepsilon\}\right)_n$ increases to $\bigcup_n \bigcap_{m \geq n} \{|f_m| \leq \varepsilon\}$. So by countable additivity,

$$\begin{aligned}\mu\left(\bigcap_{m \geq n} \{|f_m| \leq \varepsilon\}\right) &\to \mu\left(\bigcup_n \bigcap_{m \geq n} \{|f_m| \leq \varepsilon\}\right) \\ &= \mu(|f_n| \leq \varepsilon \text{ eventually}) \\ &\geq \mu(|f_n| \to 0) = \mu(E)\end{aligned}$$

Hence,

$$\liminf_n \mu(|f_n| \leq \varepsilon) \geq \mu(E) \implies \limsup_n \mu(|f_n| > \varepsilon) \leq 0 \implies \mu(|f_n| > \varepsilon) \to 0$$

For the second part, by hypothesis, we have

$$\mu\left(|f_n| > \frac{1}{k}\right) < \varepsilon$$

for sufficiently large n. So choosing $\varepsilon = \frac{1}{k^2}$, we see that along some subsequence n_k we have

$$\mu\left(|f_{n_k}| > \frac{1}{k}\right) \leq \frac{1}{k^2}$$

Hence,

$$\sum_k \mu\left(|f_{n_k}| > \frac{1}{n}\right) < \infty$$

So by the first Borel–Cantelli lemma, we have

$$\mu\left(|f_{n_k}| > \frac{1}{k} \text{ infinitely often}\right) = 0$$

so $f_{n_k} \to 0$ almost everywhere. □

Remark. Condition (i) is false if $\mu(E)$ is infinite: consider $f_n = \mathbb{1}_{(n,\infty)}$ on $(\mathbb{R}, \mathcal{B}, \mu)$, since $f_n \to 0$ almost everywhere but $\mu(f_n) = \infty$. Condition (ii) is false if we do not restrict to subsequences: consider independent events A_n such that $\mathbb{P}(A_n) = \frac{1}{n}$, then $\mathbb{1}_{A_n} \to 0$ in probability since $\mathbb{P}(\mathbb{1}_{A_n} > \varepsilon) = \mathbb{P}(A_n) = \frac{1}{n} \to 0$, but $\sum_n \mathbb{P}(A_n) = \infty$, and by the second Borel–Cantelli lemma, $\mathbb{P}(\mathbb{1}_{A_n} > \varepsilon \text{ infinitely often}) = 1$, so $\mathbb{1}_{A_n} \not\to 0$ almost surely.

Example. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent exponential random variables distributed by $\mathbb{P}(X_1 \leq x) = 1 - e^{-x}$ for $x \geq 0$. Define $A_n = \{X_n \geq \alpha \log n\}$ where $\alpha > 0$, so $\mathbb{P}(A_n) = n^{-\alpha}$, and in particular, $\sum_n \mathbb{P}(A_n) < \infty$ if and only if $\alpha > 1$. By the Borel–Cantelli lemmas, we have for all $\varepsilon > 0$,

$$\mathbb{P}\left(\frac{X_n}{\log n} \geq 1 \text{ infinitely often}\right) = 1; \quad \mathbb{P}\left(\frac{X_n}{\log n} \geq 1 + \varepsilon \text{ infinitely often}\right) = 0$$

In other words, $\limsup_n \frac{X_n}{\log n} = 1$ almost surely.

2.7. Kolmogorov's zero-one law

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables. We can define $\mathcal{T}_n = \sigma(X_{n+1}, X_{n+2}, \dots)$. Let $\mathcal{T} = \bigcap_{n\in\mathbb{N}} \mathcal{T}_n$ be the *tail σ-algebra*, which contains all events in $\mathcal{F}$ that depend only on the limiting behaviour of (X_n).

Theorem. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables. Let $A \in \mathcal{T}$ be an event in the tail σ-algebra. Then $\mathbb{P}(A) = 1$ or $\mathbb{P}(A) = 0$. If $Y : (\Omega, \mathcal{T}) \to (\mathbb{R}, \mathcal{B})$ is measurable, it is constant almost surely.

Proof. Define $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$ to be the σ-algebra generated by the first n elements of (X_n). This is also generated by the π-system of sets $A = (X_1 \leq x_1, \dots, X_n \leq x_n)$ for any $x_i \in \mathbb{R}$. Note that the π-system of sets $B = (X_{n+1} \leq x_{n+1}, \dots, X_{n+k} \leq x_{n+k})$, for arbitrary $k \in \mathbb{N}$ and $x_i \in \mathbb{R}$, generates $\mathcal{T}_n$. By independence of the sequence, we see that $\mathbb{P}(A \cap B) =$

$\mathbb{P}(A)\,\mathbb{P}(B)$ for all such sets A, B, and so the σ-algebras $\mathcal{T}_n, \mathcal{F}_n$ generated by these π-systems are independent.

Let $\mathcal{F}_\infty = \sigma(X_1, X_2, \dots)$. Then, $\bigcup_n \mathcal{F}_n$ is a π-system that generates $\mathcal{F}_\infty$. If $A \in \bigcup_n \mathcal{F}_n$, we have $A \in \mathcal{F}_n$ for some n, so there exists $\overline{n}$ such that $B \in \mathcal{T}_{\overline{n}}$ is independent of A. In particular, $B \in \bigcap_n \mathcal{T}_n = \mathcal{T}$. By uniqueness, $\mathcal{F}_\infty$ is independent of $\mathcal{T}$.

Since $\mathcal{T} \subseteq \mathcal{F}_\infty$, if $A \in \mathcal{T}$, A is independent from A. So $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)\,\mathbb{P}(A)$, so $\mathbb{P}(A)^2 - \mathbb{P}(A) = 0$ as required.

Finally, if $Y : (\Omega, \mathcal{T}) \to (\mathbb{R}, \mathcal{B})$, the preimages of $\{Y \leq y\}$ lie in $\mathcal{T}$, which give probability one or zero. Let $c = \inf\{y \mid F_Y(y) = 1\}$, so $Y = c$ almost surely. □

3. Integration

3.1. Notation

Let $f\colon (E,\mathcal{E},\mu) \to \mathbb{R}$ be an 'integrable' function, a notion we will define. We will then define the integral with respect to μ, either written $\mu(f)$ or $\int_E f \,\mathrm{d}\mu = \int_E f(x)\,\mathrm{d}\mu(x)$. If X is a random variable, we will define its expectation $\mathbb{E}\left[X\right] = \int_\Omega X \,\mathrm{d}\mathbb{P} = \int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}(\omega)$.

3.2. Definition

We say that a function $f\colon (E,\mathcal{E},\mu) \to \mathbb{R}$ is *simple* if it is of the form

$$f = \sum_{k=1}^m a_k \mathbb{1}_{A_k}; \quad a_k \geq 0; \quad A_k \in \mathcal{E}; \quad m \in \mathbb{N}$$

Definition. The *μ-integral* of a simple function f defined as above is

$$\mu(f) = \sum_{k=1}^m a_k \mu(A_k)$$

which is independent of the choice of representation of the simple function.

Remark. We have $\mu(\alpha f + \beta g) = \alpha\mu(f) + \beta\mu(g)$ for all nonnegative coefficients α, β and simple functions f, g. If $g \leq f$, $\mu(g) \leq \mu(f)$, so μ is increasing. If $f = 0$ almost everywhere, $\mu(f) = 0$.

For a general non-negative function $f\colon (E,\mathcal{E},\mu) \to \mathbb{R}$, we define its μ-integral to be

$$\mu(f) = \sup\left\{\mu(g) \mid g \leq f, g \text{ simple}\right\}$$

which agrees with the above definition for simple functions. This operator takes values in the extended non-negative real line $[0,\infty]$. Now, for $f\colon (E,\mathcal{E},\mu) \to \mathbb{R}$ measurable but not necessarily non-negative, we define $f^+ = \max(f,0)$ and $f^- = \max(-f,0)$, so that $f = f^+ - f^-$ and $|f| = f^+ + f^-$.

Definition. A measurable function $f\colon (E,\mathcal{E},\mu) \to \mathbb{R}$ is *μ-integrable* if $\mu(|f|) < \infty$. In this case, we define its integral to be

$$\mu(f) = \mu(f^+) - \mu(f^-)$$

which is a well-defined real number.

3.3. Monotone convergence theorem

Theorem. Let $f_n, f\colon (E,\mathcal{E},\mu) \to \mathbb{R}$ be measurable and non-negative such that f_n increases pointwise to f, so $f_n(x) \leq f_{n+1}(x) \leq f(x)$ and $f_n(x) \to f(x)$ as $n \to \infty$. Then, $\mu(f_n) \to \mu(f)$ as $n \to \infty$.

Remark. This is a theorem that allows us to interchange a pair of limits, $\mu(f) = \mu(\lim_n f_n) = \lim_n \mu(f_n)$. Also, $g_n \geq 0$, $\mu(\sum_n g_n) = \sum_n \mu(g_n)$.

If we consider the approximating sequence $\widetilde{f_n} = 2^{-n}\lfloor 2^n f \rfloor$, as defined in the monotone class theorem, then this is a non-negative sequence converging to f. So in particular, $\mu(f)$ is equal to the limit of the integrals of these simple functions.

It suffices to require convergence of $f_n \to f$ almost everywhere, the general argument does not need to change. The non-negativity constraint is not required if the first term in the sequence f_0 is integrable, by subtracting f_0 from every term.

Proof. Recall that $\mu(f) = \sup\{\mu(g) \mid g \leq f, g \text{ simple}\}$. Since f_n is an increasing sequence of nonnegative functions, $\mu(f_n)$ is an increasing sequence of nonnegative functions. So it converges to its (*extended* non-negative real) supremum $M = \sup_n \mu(f_n)$. Since $f_n \leq f$, $\mu(f_n) \leq \mu(f)$, so taking suprema, $M \leq \mu(f)$. If M is finite, $\sup_n \mu(f_n) = \lim_n \mu(f_n) \leq \mu(f)$. If M is infinite, we are already done.

Now, we need to show $\mu(f) \leq M$, or equivalently, $\mu(g) \leq M$ for all simple g such that $g \leq f$, so that taking suprema, $\mu(f) = \sup_g \mu(g) \leq M$. We define $g_n = \min(\overline{f}_n, g)$, where $\overline{f}_n$ is the nth approximation of f_n by simple functions from the monotone class theorem. Now, since f_n increases to f, $\overline{f}_n$ increases to f. In particular, $g_n = \min(\overline{f}_n, g)$ increases to $\min(f, g) = g$. Since $\overline{f}_n \leq f_n$ by definition, we have $g_n \leq f_n$ for all n.

Now let g be an arbitrary simple function of the form $g = \sum_{k=1}^m a_k \mathbb{1}_{A_k}$ where $a_k \geq 0$ and the $A_k \in \mathcal{E}$ are disjoint. For $\varepsilon > 0$, we define sets $A_k(n) = \{x \in A_k \mid g_n(x) \geq (1-\varepsilon)a_k\}$. Since $g = a_k$ on A_k, and since g_n increases to g, we must have $A_k(n)$ increases to A_k for all k. Since μ is a measure, $\mu(A_k(n))$ increases to $\mu(A_k)$ by countable additivity.

We have $g_n \mathbb{1}_{A_k} \geq g_n \mathbb{1}_{A_k(n)} \geq (1-\varepsilon)a_k \mathbb{1}_{A_k(n)}$ on E. Moreover, $g_n = \sum_{k=1}^m g_n \mathbb{1}_{A_k}$ since the A_k are disjoint and support g_n. Hence, $g_n \geq \sum_{k=1}^m (1-\varepsilon)a_k \mathbb{1}_{A_k(n)}$, and in particular, $\mu(g_n) \geq (1-\varepsilon)\sum_{k=1}^m a_k \mu(A_k(n))$. The right hand side increases to $(1-\varepsilon)\sum_{k=1}^m a_k \mu(A_k) = (1-\varepsilon)\mu(g)$. Hence

$$\mu(g) \leq \frac{1}{1-\varepsilon} \limsup_n \mu(g_n) \leq \frac{1}{1-\varepsilon} \limsup_n \mu(f_n) \leq \frac{M}{1-\varepsilon}$$

Since ε was arbitrary, this completes the proof. □

3.4. Linearity of integral

Theorem. Let $f, g \colon (E, \mathcal{E}, \mu) \to \mathbb{R}$ be nonnegative measurable functions. Then $\mu(\alpha f + \beta g) = \alpha\mu(f) + \beta\mu(g)$ for all $\alpha, \beta \geq 0$. Further, if $g \leq f$, then $\mu(g) \leq \mu(f)$. Finally, $f = 0$ almost everywhere if and only if $\mu(f) = 0$.

Proof. If $\widetilde{f_n}, \tilde{g}_n$ are the approximations of f and g by simple funtions from the monotone class theorem, $\alpha\widetilde{f_n}$ increases to αf and $\beta\tilde{g}_n$ increases to βg, so $\alpha\widetilde{f_n} + \beta\tilde{g}_n$ increases to $\alpha f +$

βg. Integrating both sides and using the monotone convergence theorem, the result follows, since linearity of simple functions is simple to prove.

The second part $g \leq f \implies \mu(g) \leq \mu(f)$ has already been proven. Now, if $f = 0$ almost everywhere, its approximation $0 \leq \widetilde{f_n}$ increases to f almost everywhere, so must be exactly zero for all n. So $\mu(\widetilde{f_n}) = 0$ so $\mu(f) = 0$. Conversely, if $\mu(f) = 0$, then $0 \leq \mu(\widetilde{f_n}) \to 0$ gives $\mu(\widetilde{f_n}) = 0$ so $\widetilde{f_n} = 0$ almost everywhere. Since $0 = \widetilde{f_n}$ increases almost everywhere to f, f is zero almost everywhere. □

Remark. Functions such as $\mathbb{1}_{\mathbb{Q}}$ are integrable and have integral zero. They are 'identified' with the zero element in the theory of integration.

Theorem. Let $f, g : (E, \mathcal{E}, \mu) \to \mathbb{R}$ be integrable functions. Then $\mu(\alpha f + \beta g) = \alpha\mu(f) + \beta\mu(g)$ for all $\alpha, \beta \in \mathbb{R}$; if $g \leq f$, then $\mu(g) \leq \mu(f)$; and if $f = 0$ almost everywhere, we have $\mu(f) = 0$.

Proof. Clearly, if f is integrable, so is αf, and $\mu(-f) = -\mu(f)$, by definition of the integral for a general function. We can explicitly check that for $\alpha \geq 0$, we have $\mu(\alpha f) = \mu((\alpha f)^+) - \mu((\alpha f)^-) = \alpha\mu(f^+) - \alpha\mu(f^-) = \alpha\mu(f)$. Define $h = f + g$. Then $h^+ + f^- + g^- = h^- + f^+ + g^+$, so by the previous theorem, $\mu(h^+) + \mu(f^-) + \mu(g^-) = \mu(h^-) + \mu(f^+) + \mu(g^+)$ and the result holds.

Finally, if $0 \leq f - g$, we have $0 \leq \mu(0) \leq \mu(f - g) = \mu(f) - \mu(g)$ so the result follows. If $f = 0$ almost everywhere, $f^+ = 0$ and $f^- = 0$ almost everywhere, so $\mu(f) = 0$. □

3.5. Fatou's lemma

Lemma. Let $f_n : (E, \mathcal{E}, \mu) \to \mathbb{R}$ be nonnegative measurable functions. Then $\mu(\liminf_n f_n) \leq \liminf_n \mu(f_n)$.

Remark. Recall that $\liminf_n x_n = \sup_n \inf_{m \geq n} x_m$ and $\limsup_n x_n = \inf_n \sup_{m \geq n} x_m$. In particular, $\limsup_n x_n = \liminf_n x_n$ implies that $\lim_n x_n$ exists and is equal to $\limsup_n x_n$ and $\liminf_n x_n$. Hence, if the f_n converge to some measurable function f, we must have $\mu(f) \leq \liminf_n \mu(f_n)$.

Proof. We have $\inf_{m \geq n} f_m \leq f_k$ for all $k \geq n$, so by taking integrals, $\mu\big(\inf_{m \geq n} f_m\big) \leq \mu(f_k)$. Thus,

$$\mu\left(\inf_{m \geq n} f_m\right) \leq \inf_{k \geq n} \mu(f_k) \leq \sup_n \inf_{k \geq n} \mu(f_k) = \liminf_n \mu(f_n)$$

Note that $\inf_{m \geq n} f_m$ increases to $\sup_n \inf_{m \geq n} f_m = \liminf_n f_n$. By the monotone convergence theorem,

$$\mu\left(\liminf_n f_n\right) = \lim_n \mu\left(\inf_{m \geq n} f_m\right) \leq \liminf_n \mu(f_n)$$

as required. □

3.6. Dominated convergence theorem

Theorem. Let $f_n, f : (E, \mathcal{E}, \mu)$ be measurable functions such that $|f_n| \leq g$ almost everywhere on E, and the dominating function g is μ-integrable, so $\mu(g) < \infty$. Suppose $f_n \to f$ pointwise (or almost everywhere) on E. Then f_n and f are also integrable, and $\mu(f_n) \to \mu(f)$ as $n \to \infty$.

Proof. Clearly $\mu(|f_n|) \leq \mu(g) < \infty$, so the f_n are integrable. Taking limits in $|f_n| \leq g$, we have $|f| \leq g$, so f is also integrable by the same argument. Now, $g \pm f_n$ is a nonnegative function, and converges pointwise to $g \pm f$. Since limits are equal to the limit inferior when they exist, by Fatou's lemma, we have

$$\mu(g) + \mu(f) = \mu(g + f) = \mu\left(\liminf_n (g + f_n)\right) \leq \liminf_n \mu(g + f_n) = \mu(g) + \liminf_n \mu(f_n)$$

Hence $\mu(f) \leq \liminf_n \mu(f_n)$. Likewise, $\mu(g) - \mu(f) \leq \mu(g) - \liminf_n \mu(f_n)$, so $\mu(f) \geq \limsup_n \mu(f_n)$, so

$$\limsup_n \mu(f_n) \leq \mu(f) \leq \liminf_n \mu(f_n)$$

But since $\liminf_n \mu(f_n) \leq \limsup_n \mu(f_n)$, the result follows. □

Example. Let $E = [0,1]$ with the Lebesgue measure. Let $f_n \to f$ pointwise and the f_n are uniformly bounded, so $\sup_n \|f_n\|_\infty \leq g$ for some $g \in \mathbb{R}$. Then since $\mu(g) = g < \infty$, the dominated convergence theorem implies that f_n, f are integrable and $\mu(f_n) \to \mu(f)$ as $n \to \infty$. In particular, no notion of uniform convergence of the f_n is required.

Remark. The proof of the fundamental theorem of calculus requires only the fact that

$$\int_x^{x+h} \mathrm{d}t = h$$

This is a fact which is obviously true of the Riemann integral and also of the Lebesgue integral. Therefore, for any continuous function $f : [0,1] \to \mathbb{R}$, we have

$$\underbrace{\int_0^x f(t)\,\mathrm{d}t}_{\text{Riemann integral}} = F(x) = \underbrace{\int_0^x f(t)\,\mathrm{d}\mu(t)}_{\text{Lebesgue integral}}$$

So these integrals coincide for continuous functions. We can show that all Riemann integrable functions are $\mu^\star$-measurable, where $\mu^\star$ is the outer measure of the Lebesgue measure, as defined in the proof of Carathéodory's theorem. However, there exist certain Riemann integrable functions that are not Borel measurable. We can find that a bounded $\mu^\star$-measurable function is Riemann integrable if and only if

$$\mu(\{x \in [0,1] \mid f \text{ is discontinuous at } x\}) = 0$$

The standard techniques of Riemann integration, such as substitution and integration by parts, extend to all bounded measurable functions by the monotone class theorem.

Theorem. Let $U \subseteq \mathbb{R}$ be an open set and $(E, \mathcal{E}, \mu)$ be a measure space. Let $f : U \times E \to \mathbb{R}$ be a map such that $x \mapsto f(t, x)$ is measurable, and $t \mapsto f(t, x)$ is differentiable where $\left|\frac{\partial f}{\partial t}\right| < g(x)$ for all $t \in U$, and g is μ-integrable. Then

$$F(t) = \int_E f(t,x)\,\mathrm{d}\mu(x) \implies F'(t) = \int_E \frac{\partial f}{\partial t}(t,x)\,\mathrm{d}\mu(x)$$

Proof. By the mean value theorem,

$$g_h(x) = \frac{f(t+h,x) - f(t,x)}{h} - \frac{\partial f}{\partial t}(t,x) \implies |g_h(x)| = \left|\frac{\partial f}{\partial t}(\tilde{t},x) - \frac{\partial f}{\partial t}(t,x)\right| \leq 2g(x)$$

Note that g is μ-integrable. By differentiability of f, we have $g_h \to 0$ as $h \to 0$, so applying the dominated convergence theorem, $\mu(g_h) \to \mu(0) = 0$. By linearity of the integral,

$$\mu(g_h) = \frac{\int_E f(t+h,x) - f(t,x)\,\mathrm{d}\mu(x)}{h} - \int_E \frac{\partial f}{\partial t}(t,x)\,\mathrm{d}\mu(x)$$

Hence, $\frac{F(t+h)-F(t)}{h} - F'(t) \to 0$. □

Example. For a measurable function $f : (E, \mathcal{E}, \mu) \to (G, \mathcal{G})$, if $g : G \to \mathbb{R}$ is a nonnegative function, we show on an example sheet that

$$\mu \circ f^{-1}(g) = \int_G g\,\mathrm{d}\mu \circ f^{-1} = \int_E g(f(x))\,\mathrm{d}\mu(x) = \mu(g \circ f)$$

On a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a G-valued random variable X, we then compute

$$\mathbb{E}\left[g(X)\right] = \mu_X(g) = \int_\Omega g(X(\omega))\,\mathrm{d}\mathbb{P}(\omega) = \int_\Omega g\,\mathrm{d}\mathbb{P}$$

Example (measures with densities)**.** If $f : (E, \mathcal{E}, \mu) \to \mathbb{R}$ is a nonnegative measurable function, we can define $\nu_f(A) = \mu(f 1_A)$ for any measurable set A, which is again a measure on $(E, \mathcal{E})$ by the monotone convergence theorem. In particular, if $g : (E, \mathcal{E}) \to \mathbb{R}$ is measurable, $\nu_f(g) = \int_E g(x) f(x)\,\mathrm{d}\mu(x) = \int_E g\,\mathrm{d}\nu(f)$. We call f the *density* of ν_f with respect to μ. If its integral is one, it is called a *probability density function*.

4. Product measures

4.1. Integration in product spaces

Let $(E_1, \mathcal{E}_1, \mu_1), (E_2, \mathcal{E}_2, \mu_2)$ be finite measure spaces. On $E = E_1 \times E_2$, we can consider the π-system of 'rectangles' $\mathcal{A} = \{A_1 \times A_2 \mid A_1 \in \mathcal{E}_1, A_2 \in \mathcal{E}_2\}$. Then we define the σ-algebra $\mathcal{E}_1 \otimes \mathcal{E}_2 = \sigma(\mathcal{A})$ on the product space. If the E_i are topological spaces with a countable base, then $\mathcal{B}(E_1 \times E_2) = \mathcal{B}(E_1) \otimes \mathcal{B}(E_2)$.

Lemma. Let $E = E_1 \times E_2, \mathcal{E} = \mathcal{E}_1 \otimes \mathcal{E}_2$. Let $f : (E, \mathcal{E}) \to \mathbb{R}$ be measurable. Then for all $x_1 \in E_1$, the map $(x_2 \mapsto f(x_1, x_2)) : (E_2, \mathcal{E}_2) \to \mathbb{R}$ is $\mathcal{E}_2$-measurable.

Proof. Let

$$\mathcal{V} = \{f : (E, \mathcal{E}) \to \mathbb{R} \mid f \text{ bounded, measurable, conclusion of the lemma holds}\}$$

This is a $\mathbb{R}$-vector space, and it contains $\mathbb{1}_E, \mathbb{1}_A$ for all $A \in \mathcal{A}$, since $\mathbb{1}_A = \mathbb{1}_{A_1(x_1)}\mathbb{1}_{A_2(x_2)}$. Now, let $0 \leq f_n$ increase to f, $f_n \in \mathcal{V}$. Then $(x_2 \mapsto f(x_1, x_2)) = \lim_n (x_2 \mapsto f_n(x_1, x_2))$, so it is $\mathcal{E}_2$-measurable as a limit of a sequence of measurable functions. Then by the monotone class theorem, $\mathcal{V}$ contains all bounded measurable functions. This extends to all measurable functions by truncating the absolute value of f to $n \in \mathbb{N}$, then the sequence of such bounded truncations converges pointwise to f. □

Lemma. Let $E = E_1 \times E_2, \mathcal{E} = \mathcal{E}_1 \otimes \mathcal{E}_2$. Let $f : (E, \mathcal{E}) \to \mathbb{R}$ be measurable such that

(i) f is bounded; or

(ii) f is nonnegative.

Then the map $x_1 \mapsto \int_{E_2} f(x_1, x_2) \, \mathrm{d}\mu_2(x_2)$ is μ_1-measurable and is bounded or nonnegative respectively.

Remark. In case (ii), the map on x_1 may evaluate to infinity, but the set of values

$$\left\{ x_1 \in E_1 \,\middle|\, \int_{E_2} f(x_1, x_2) \, \mathrm{d}\mu_2(x_2) = \infty \right\}$$

lies in $\mathcal{E}_1$.

Proof. Let

$$\mathcal{V} = \{f : (E, \mathcal{E}) \to \mathbb{R} \mid f \text{ bounded, measurable, conclusion of the lemma holds}\}$$

This is a vector space by linearity of the integral. $\mathbb{1}_E \in \mathcal{V}$, since $\mathbb{1}_{E_1}\mu_2(E_2)$ is nonnegative and bounded. $\mathbb{1}_A \in \mathcal{V}$ for all $A \in \mathcal{A}$, because $\mathbb{1}_{A_1}(x_1)\mu_2(A_2)$ is $\mathcal{E}_1$-measurable, nonnegative, and bounded since it is at most $\mu_2(E_2) < \infty$. Now let f_n be a sequence of nonnegative functions that increase to f, where $f_n \in \mathcal{V}$. Then by the monotone convergence theorem,

$$\int_{E_2} \lim_{n\to\infty} f_n(x_1, x_2) \, \mathrm{d}\mu_2(x_2) = \lim_{n\to\infty} \int_{E_2} f_n(x_1, x_2) \, \mathrm{d}\mu_2(x_2)$$

is an increasing limit of $\mathcal{E}_1$-measurable functions, so is $\mathcal{E}_1$-measurable. It is bounded by $\mu_2(E_2)\|f\|_\infty$, or nonnegative as required. So $f \in \mathcal{V}$. By the monotone class theorem, the result for bounded functions holds. In case (ii), we can take a bounded approximation in $\mathcal{V}$ of an arbitrary measurable function f to conclude the proof. □

Theorem (product measure). Let $(E_1, \mathcal{E}_1, \mu_1), (E_2, \mathcal{E}_2, \mu_2)$ be finite measure spaces. There exists a unique measure $\mu = \mu_1 \otimes \mu_2$ on $(E_1 \times E_2, \mathcal{E}_1 \otimes \mathcal{E}_2)$ such that $\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$ for all $A_1 \in \mathcal{E}_1, A_2 \in \mathcal{E}_2$.

Proof. $\mathcal{A}$ generates $\mathcal{E} \otimes \mathcal{E}_2$, so by the uniqueness theorem, there can only be one such measure. We define

$$\mu(A) = \int_{E_1} \left(\int_{E_2} \mathbb{1}_A(x_1, x_2) \, \mathrm{d}\mu_2(x_2) \right) \mathrm{d}\mu_1(x_1)$$

We have

$$\begin{aligned} \mu(A_1 \times A_2) &= \int_{E_1} \left(\int_{E_2} \mathbb{1}_{A_1}(x_1)\mathbb{1}_{A_2}(x_2) \, \mathrm{d}\mu_2(x_2) \right) \mathrm{d}\mu_1(x_1) \\ &= \int_{E_1} \mathbb{1}_{A_1}(x_1)\mu_2(A_2) \, \mathrm{d}\mu_1(x_1) \\ &= \mu_1(A_1)\mu_2(A_2) \end{aligned}$$

Clearly $\mu(\varnothing) = 0$, so it suffices to show countable additivity. Let A_n be disjoint sets in $\mathcal{E}_1 \otimes \mathcal{E}_2$. Then

$$\mathbb{1}_{(\bigcup_n A_n)} = \sum_n \mathbb{1}_{A_n} = \lim_{n\to\infty} \sum_{i=1}^n \mathbb{1}_{A_n}$$

Then by the monotone convergence theorem and the previous lemmas,

$$\begin{aligned} \mu\left(\bigcup_n A_n\right) &= \int_{E_1} \left(\int_{E_2} \lim_{n\to\infty} \sum_{i=1}^n \mathbb{1}_{A_i} \, \mathrm{d}\mu_2(x_2) \right) \mathrm{d}\mu_1(x_1) \\ &= \int_{E_1} \left(\lim_{n\to\infty} \int_{E_2} \sum_{i=1}^n \mathbb{1}_{A_i} \, \mathrm{d}\mu_2(x_2) \right) \mathrm{d}\mu_1(x_1) \\ &= \lim_{n\to\infty} \int_{E_1} \left(\int_{E_2} \sum_{i=1}^n \mathbb{1}_{A_i} \, \mathrm{d}\mu_2(x_2) \right) \mathrm{d}\mu_1(x_1) \\ &= \lim_{n\to\infty} \sum_{i=1}^n \int_{E_1} \left(\int_{E_2} \mathbb{1}_{A_i} \, \mathrm{d}\mu_2(x_2) \right) \mathrm{d}\mu_1(x_1) \\ &= \lim_{n\to\infty} \sum_{i=1}^n \mu(A_i) \\ &= \sum_{n=1}^\infty \mu(A_n) \end{aligned}$$

□

4.2. Fubini's theorem

Theorem. Let $(E, \mathcal{E}, \mu) = (E_1 \times E_2, \mathcal{E}_1 \otimes \mathcal{E}_2, \mu_1 \otimes \mu_2)$ be a finite measure space. Let $f : E \to \mathbb{R}$ be a nonnegative measurable function. Then

$$\begin{aligned} \mu(f) &= \int_E f \, \mathrm{d}\mu \\ &= \int_{E_1} \left(\int_{E_2} f(x_1, x_2) \, \mathrm{d}\mu_2(x_2) \right) \mathrm{d}\mu_1(x_1) \\ &= \int_{E_2} \left(\int_{E_1} f(x_1, x_2) \, \mathrm{d}\mu_1(x_1) \right) \mathrm{d}\mu_2(x_2) \end{aligned}$$

Now, let $f : E \to \mathbb{R}$ be a μ-integrable function (on the product measure). Let

$$A_1 = \left\{ x_1 \in E_1 \,\middle|\, \int_{E_2} |f(x_1, x_2)| \, \mathrm{d}\mu_2(x_2) < \infty \right\}$$

Define f_1 by $f_1(x_1) = \int_{E_2} f(x_1, x_2) \, \mathrm{d}\mu_2(x_2)$ on A_1 and zero elsewhere. Then $\mu_1(A_1^{\mathrm{c}}) = 0$ and $\mu(f) = \mu_1(f_1) = \mu_1(f_1 \mathbb{1}_{A_1})$, and defining A_2 symmetrically, $\mu(f) = \mu_2(f_2) = \mu_2(f_2 \mathbb{1}_{A_2})$.

Remark. If f is bounded, $A_1 = E_1$. Note, for $f(x_1, x_2) = \frac{x_1^2 - x_2^2}{(x_1^2 + x_2^2)^2}$ on $(0,1)^2$, we have $\mu_1(f_1) \neq \mu_2(f_2)$, but f is not Lebesgue integrable on $(0,1)^2$.

Proof. By the construction of the product measure $\mu(A)$ for rectangles $A = A_1 \times A_2$ in the π-system $\mathcal{A}$ generating $\mathcal{E}$, the identities in the first part of the theorem clearly hold for $f = \mathbb{1}_A$. By uniqueness, this extends to $\mathbb{1}_A$ for all $A \in \mathcal{E}$. Then, by linearity of the integral, this extends to simple functions. By the monotone convergence theorem, the first part of the theorem follows.

Now let f be μ-integrable. Let $h(x_1) = \int_{E_2} |f(x_1, x_2)| \, \mathrm{d}\mu_2(x_2)$. Then by the first part, $\mu_1(|h|) \leq \mu(|f|) < \infty$. So f_1 is μ_1-integrable. We have $\mu_1(A_1^{\mathrm{c}}) = 0$, otherwise, we could compute a lower bound $\mu_1(|h|) \geq \mu_1(|h| \mathbb{1}_{A_1^{\mathrm{c}}}) = \infty$, but it must be finite. Note that $f_1^{\pm} = \int_{E_2} f^{\pm}(x_1, x_2) \, \mathrm{d}\mu_2(x_2)$, and $\mu(f_1) = \mu_1(f_1^+) - \mu_1(f_1^-)$. Hence, by the first part, $\mu(f) = \mu(f^+) - \mu(f^-) = \mu_1(f_1^+) - \mu_1(f_1^-) = \mu_1(f_1)$ as required. □

Remark. The proofs above extend to σ-finite measures μ.

Let $(E_i, \mathcal{E}_i, \mu_i)$ be measure spaces with σ-finite measures. Note that $(\mathcal{E}_1 \otimes \mathcal{E}_2) \otimes \mathcal{E}_3 = \mathcal{E}_1 \otimes (\mathcal{E}_2 \otimes \mathcal{E}_3)$, by a π-system argument using Dynkin's lemma. So we can iterate the construction of the product measure to obtain a measure $\mu_1 \otimes \dots \mu_n$, which is a unique measure on $\left(\prod_{i=1}^n E_i \bigotimes_{i=1}^n \mathcal{E}_i \right)$ with the property that the measure of a hypercube $\mu(A_1 \times A_n)$ is the product of the measures of its sides $\mu_i(A_i)$.

In particular, we have constructed the Lebesgue measure $\mu^n = \bigotimes_{i=1}^n \mu$ on $\mathbb{R}^n$. Applying Fubini's theorem, for functions f that are either nonnegative and measurable or μ^n-integrable, we have

$$\int_{\mathbb{R}^n} f \, \mathrm{d}\mu^n = \int \cdots \int_{\mathbb{R} \dots \mathbb{R}} f(x_1, \dots, x_n) \, \mathrm{d}\mu(x_1) \dots \mathrm{d}\mu(x_n)$$

4.3. Product probability spaces and independence

Proposition. Let $(\Omega, \mathcal{F}, \mathbb{P})$, and $(E, \mathcal{E}) = \left(\prod_{i=1}^n E_i, \bigotimes_{i=1}^n \mathcal{E}_i\right)$. Let $X : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ be a measurable function, and define $X(\omega) = (X_1(\omega), X_2(\omega), \dots, X_n(\omega))$. Then the following are equivalent.

(i) $X_1, \dots, X_n$ are independent random variables;

(ii) $\mu_X = \bigotimes_{i=1}^n \mu_{X_i}$;

(iii) for all bounded and measurable $f_i : E_i \to \mathbb{R}$, $\mathbb{E}\left[\prod_{i=1}^n f_i(X_i)\right] = \prod_{i=1}^n \mathbb{E}\left[f_i(X_i)\right]$.

Proof. (i) implies (ii). Consider the π-system $\mathcal{A}$ of rectangles $A = \prod_{i=1}^n A_i$ for $A_i \in \mathcal{E}_i$. Since μ_X is an image measure, Then

$$\mu_X(A_1 \times \dots \times A_n) = \mathbb{P}\left(X_1 \in A_1, \dots, X_n \in A_n\right) = \mathbb{P}\left(X_1\right) \dots \mathbb{P}\left(A_n\right) = \prod_{i=1}^n \mu_{X_i}(A_i)$$

So by uniqueness, the result follows.

(ii) implies (iii). By Fubini's theorem,

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=1}^n f_i(X_i)\right] &= \mu_X\left(\prod_{i=1}^n f_i(x_i)\right) \\
&= \int_E f(x) \, \mathrm{d}\mu(x) \\
&= \int \cdots \int_{E_i} \left(\prod_{i=1}^n f_i(x_i)\right) \mathrm{d}\mu_{X_1}(x_1) \dots \mathrm{d}\mu_{X_2}(x_2) \\
&= \prod_{i=1}^n \int_{E_i} f_i(x_i) \, \mathrm{d}\mu_{X_i}(x_i) \\
&= \prod_{i=1}^n \mathbb{E}\left[f_i(X_i)\right]
\end{aligned}$$

(iii) implies (i). Let $f_i = \mathbb{1}_{A_i}$ for any $A_i \in \mathcal{E}_i$. These are bounded and measurable functions. Then

$$\mathbb{P}\left(X_1 \in A_1, \dots, X_n \in A_n\right) = \mathbb{E}\left[\prod_{i=1}^n \mathbb{1}_{A_i}(X_i)\right] = \prod_{i=1}^n \mathbb{E}\left[\mathbb{1}_{A_i}(X_i)\right] = \prod_{i=1}^n \mathbb{P}\left(X_i \in A_i\right)$$

So the σ-algebras generated by the X_i are independent as required. □

5. Function spaces and norms

5.1. Norms

Definition. A *norm* on a real vector space is a map $\| \cdot \|_V : V \to \mathbb{R}$ such that

(i) $\|\lambda v\| = |\lambda| \cdot \|v\|$;

(ii) $\|u + v\| \leq \|u\| + \|v\|$;

(iii) $\|v\| = 0$ if and only if $v = 0$.

Definition. Let $(E, \mathcal{E}, \mu)$ be a measure space. We define $L^p(E, \mathcal{E}, \mu) = L^p(\mu) = L^p$ for the space of measurable functions $f : E \to \mathbb{R}$ such that $\|f\|_p$ is finite, where

$$\|f\|_p = \begin{cases} \left(\int_E |f(x)|^p \, \mathrm{d}\mu(x)\right)^{\frac{1}{p}} & 1 \leq p < \infty \\ \operatorname{ess\,sup} |f| = \inf\{\lambda > 0 \mid |f| \leq \lambda \text{ almost everywhere}\} & p = \infty \end{cases}$$

We must check that $\| \cdot \|_p$ as defined is a norm. Clearly (i) holds for all $1 \leq p \leq \infty$. Property (ii) holds for $p = 1$ and $p = \infty$, and we will prove later that this holds for other values of p. The last property does not hold: $f = 0$ implies $\|f\|_p = 0$, but $\|f\|_p = 0$ implies only that $|f|^p = 0$ almost everywhere, so f is zero almost everywhere on E. Therefore, to rigorously define the norm, we must construct the quotient space $\mathcal{L}^p$ of functions that coincide almost everywhere. We write $[f]$ for the equivalence class of functions that are equal almost everywhere. The functional $\| \cdot \|_p$ is then a norm on $\mathcal{L}^p$.

Proposition (Chebyshev's inequality, Markov's inequality)**.** Let $f : E \to \mathbb{R}$ be nonnegative and measurable. Then for all $\lambda > 0$,

$$\mu(\{x \in E \mid f(x) \geq \lambda\}) = \mu(f \geq \lambda) \leq \frac{\mu(f)}{\lambda}$$

Proof. Integrate the inequality $\lambda 1_{\{f \geq \lambda\}} \leq f$, which holds on E. □

Definition. Let $I \subseteq R$ be an interval. Then we say a map $c : I \to \mathbb{R}$ is *convex* if for all $x, y \in I$ and $t \in [0,1]$, we have $c(tx + (1-t)y) \leq tc(x) + (1-t)c(y)$. Equivalently, for all $x < t < y$ and $x, y \in I$, we have $\frac{c(t)-c(x)}{t-x} \leq \frac{c(y)-c(t)}{y-t}$.

Since a convex function is continuous on the interior of the interval, it is Borel measurable.

Lemma. Let $I \subseteq R$ be an interval, and let $m \in I^\circ$. If c is convex on I, there exist a, b such that $c(x) \geq ax + b$, and $c(m) = am + b$.

Proof. Define $a = \sup\left\{\frac{c(m)-c(x)}{m-x} \mid x < m, x \in I\right\}$. This exists in $\mathbb{R}$ by the second definition of convexity. Let $y \in I$, and $y > m$. Then $a \leq \frac{c(y)-c(m)}{y-m}$, so $c(y) \geq ay - am + c(m) = ay + b$

where we define $b = c(m) - am$. Similarly, for $y < m$, by definition of the supremum, $\frac{c(m)-c(y)}{m-y} \leq a$, we have $c(y) \geq ay + b$. □

Theorem (Jensen's inequality)**.** Let X be a random variable taking values in an interval $I \subseteq \mathbb{R}$, such that $\mathbb{E}\left[|X|\right] < \infty$. Let $c\colon I \to \mathbb{R}$ be a convex function. Then $c(\mathbb{E}\left[X\right]) \leq \mathbb{E}\left[c(X)\right]$.

Note that the integral $\mathbb{E}\left[c(X)\right]$ is defined as $\mathbb{E}\left[c^+(X)\right] - \mathbb{E}\left[c^-(X)\right]$, and this is well-defined and takes values in $(-\infty, \infty]$.

Proof. Define $m = \mathbb{E}\left[X\right] = \int_I z \,\mathrm{d}\mu_X(z)$. If $m \notin I^\circ$, X must equal m almost surely, and then the result follows. Now let $m \in I^\circ$. Applying the previous lemma, we find a, b such that $c^-(X) \leq |a| \cdot |X| + |b|$. Hence, $\mathbb{E}\left[c^-(X)\right] \leq |a|\mathbb{E}\left[|X|\right] + |b| < \infty$, and $\mathbb{E}\left[c(X)\right] = \mathbb{E}\left[c^+(X)\right] - \mathbb{E}\left[c^-(X)\right]$ is well-defined in $(-\infty, \infty]$. Integrating the inequality from the lemma, and using linearity of the integral,

$$\mathbb{E}\left[c(X)\right] \geq a\mathbb{E}\left[X\right] + b = am + b = c(m) = c(\mathbb{E}\left[X\right])$$

□

Remark. If $1 \leq p < q < \infty$, $c(x) = |x|^{\frac{q}{p}}$ is a convex function. If X is a bounded random variable (so lies in $L^\infty(\mathbb{P})$), we then have

$$\|X\|_p = \mathbb{E}\left[|X^p|\right]^{\frac{1}{p}} = c(\mathbb{E}\left[|X|^p\right])^{\frac{1}{q}} \leq \mathbb{E}\left[c(|X|^p)\right]^{\frac{1}{q}} = \|X\|_q$$

Using the monotone convergence theorem, this extends to all $X \in L^q(\mathbb{P})$ when $\|X\|_q$ is finite. In particular, $L^q(\mathbb{P}) \subseteq L^p(\mathbb{P})$ for all $1 \leq p \leq q \leq \infty$.

Theorem (Hölder's inequality)**.** Let f, g be measurable functions on $(E, \mathcal{E}, \mu)$. If p, q are *conjugate*, so $\frac{1}{p} + \frac{1}{q} = 1$ and $1 \leq p, q \leq \infty$, we have

$$\mu(|fg|) = \int_E |f(x)g(x)| \,\mathrm{d}\mu \leq \|f\|_p \cdot \|g\|_q$$

Remark. For $p = q = 2$, this is exactly the Cauchy–Schwarz inequality on L^2.

Proof. The cases $p = 1$ or $p = \infty$ are obvious. We can assume $f \in L^p$ and $g \in L^q$ without loss of generality since the right hand side would otherwise be infinite. We can also assume f is not equal to zero almost everywhere, otherwise this reduces to $0 \leq 0$. Hence, $\|f\|_p > 0$. Then, we can divide both sides by $\|f\|_p$ and then assume $\|f\|_p = 1$.

$$\mu(|fg|) = \int_E |g| \frac{1}{|f|^{p-1}} |f|^p \mathbb{1}_{\{|f|>0\}} \,\mathrm{d}\mu$$

Note that we can set $|f|^p \,\mathrm{d}\mu = \mathrm{d}\mathbb{P}$, and since $L^q(\mathbb{P}) \subseteq L^1(\mathbb{P})$,

$$\int_E |g| \frac{1}{|f|^{p-1}} |f|^p \mathbb{1}_{\{|f|>0\}} \,\mathrm{d}\mu \leq \left(\int |g|^q \frac{1}{|f|^{q(p-1)}} \underbrace{|f|^p \,\mathrm{d}\mu}_{\mathrm{d}\mathbb{P}} \right)^{\frac{1}{q}} = \left(\int_E |g|^q \,\mathrm{d}\mu \right)^{\frac{1}{q}}$$

□

Theorem (Minkowski's inequality)**.** Let $f, g : (E, \mathcal{E}, \mu) \to \mathbb{R}$ be measurable functions. Then for all $1 \leq p \leq \infty$, we have $\|f + g\|_p \leq \|f\|_p + \|g\|_p$.

Proof. The results for $p = 1, \infty$ are clear. Suppose $1 < p < \infty$. We can assume without loss of generality that $f, g \in L^p$. We can integrate the pointwise inequality $|f + g|^p \leq 2^p(|f|^p + |g|^p)$ to deduce that $\|f + g\|_p^p \leq 2^p(\|f\|_p^p + \|g\|_p^p) < \infty$ so $f + g \in L^p$. We assume that $0 < \|f + g\|_p$, otherwise the result is trivial. Now, using Hölder's inequality with q conjugate to p,

$$\begin{aligned} \|f + g\|_p^p &= \int_E |f + g|^{p-1} |f + g| \,\mathrm{d}\mu \\ &\leq \int_E |f + g|^{p-1} |f| \,\mathrm{d}\mu + \int_E |f + g|^{p-1} |g| \,\mathrm{d}\mu \\ &\leq \left(\int_E |f + g|^{q(p-1)} \,\mathrm{d}\mu \right)^{\frac{1}{q}} \left(\|f\|_p + \|g\|_p \right) \\ &\leq \left(\int_E |f + g|^p \,\mathrm{d}\mu \right)^{\frac{1}{q}} \left(\|f\|_p + \|g\|_p \right) \\ &\leq \|f + g\|_p^{\frac{p}{q}} \left(\|f\|_p + \|g\|_p \right) \end{aligned}$$

Dividing both sides by $\|f + g\|_p^{\frac{p}{q}}$, we obtain $\|f + g\|_p \leq \|f\|_p + \|g\|_p$. □

So the L^p spaces are indeed normed spaces.

5.2. Banach spaces

Definition. A *Banach space* is a complete normed vector space.

Theorem ($\mathcal{L}^p$ is a Banach space)**.** Let $1 \leq p \leq \infty$, and let $f_n \in L^p$ be a Cauchy sequence, so for all $\varepsilon > 0$ there exists N such that for all $m, n \geq N$, we have $\|f_m - f_n\|_p < \varepsilon$. Then there exists a function $f \in L^p$ such that $f_n \to f$ in L^p, so $\|f_n - f\|_p \to 0$ as $n \to \infty$.

Proof. For this proof, we assume $p < \infty$; the other case is already proven in IB Analysis and Topology. Since f_n is Cauchy, using $\varepsilon = 2^{-k}$ we extract a subsequence f_{N_k} of L^p functions such that

$$S = \sum_{k=1}^{\infty} \left\| f_{N_{k+1}} - f_{N_k} \right\|_p \leq \sum_{k=1}^{\infty} 2^{-k} < \infty$$

By Minkowski's inequality, for any K, we have

$$\left\| \sum_{k=1}^{K} \left| f_{N_{k+1}} - f_{N_k} \right| \right\|_p \leq \sum_{k=1}^{K} \left\| f_{N_{k+1}} - f_{N_k} \right\|_p \leq S < \infty$$

By the monotone convergence theorem applied to $\left| \sum_{k=1}^{K} \left| f_{N_{k+1}} - f_{N_k} \right| \right|^p$ which increases to $\left| \sum_{k=1}^{\infty} \left| f_{N_{k+1}} - f_{N_k} \right| \right|^p$, we find

$$\left\| \sum_{k=1}^{\infty} \left| f_{N_{k+1}} - f_{N_k} \right| \right\|_p \leq S < \infty$$

Since the integral is finite, we see that $\sum_{k=1}^{\infty} \left| f_{N_{k+1}} - f_{N_k} \right|$ is finite almost everywhere. Then $\sum_{k=1}^{K} (f_{N_{k+1}}(x) - f_{N_k}(x)) = f_{N_{k+1}}(x) - f_{N_1}(x)$ converges in the real line for all x in a set A that has full measure, so $\mu(A^c) = 0$. In particular, $f_{N_k}(x)$ is a Cauchy sequence of reals, so by completeness of the real line, we can define the limit

$$f(x) = \begin{cases} \lim_{k \to \infty} f_{N_k}(x) & x \in A \\ 0 & x \in A^c \end{cases}$$

so $f_{N_k} \to f$ as $k \to \infty$ almost everywhere. Now, by Fatou's lemma,

$$\|f_n - f\|_p^p = \mu(|f_n - f|^p) = \mu(\lim_k \left| f_n - f_{N_k} \right|^p) \leq \liminf_k \mu(\left| f_n - f_{N_k} \right|^p)$$

Since the f_n are Cauchy,

$$\|f\|_p \leq \underbrace{\|f - f_N\|_p}_{\leq \varepsilon} + \underbrace{\|f_N\|_p}_{<\infty} < \infty$$

so $f \in L^p$, and $\|f_n - f\|_p^p \leq \varepsilon^p$ for $n, N_k \geq N$, so $f_n \to f$ in L^p. □

Remark. If V is any of the spaces

$$C([a,b]); \quad \{f \text{ simple}\}; \quad \{f \text{ a linear combination of indicators of intervals}\}$$

then V is dense in $L^1(\mu)$ where μ is the Lebesgue measure on $\mathcal{B}([a,b])$. So the completion $\overline{(V, \|\cdot\|)}$ is exactly $L^1(\mu)$.

5.3. Hilbert spaces

Definition. A symmetric bilinear form $\langle \cdot , \cdot \rangle : V \times V \to \mathbb{R}$ on a real vector space V is called an *inner product* if $\langle v, v\rangle \geq 0$ and $\langle v, v\rangle = 0$ implies $v = 0$. In this case, we can define a norm $\|v\| = \sqrt{\langle v, v\rangle}$. If $(V, \langle \cdot , \cdot \rangle)$ is complete, we say that it is a *Hilbert space*.

Corollary. The space $\mathcal{L}^2$ is a Hilbert space for the inner product $\langle f, g\rangle = \int_E fg \, \mathrm{d}\mu$.

Example. An analog of the Pythagorean theorem holds. Let $f, g \in L^2$, then $\|f+g\|_2^2 = \|f\|_2^2 + 2\langle f, g\rangle + \|g\|_2^2$. We say f is *orthogonal* to g if $\langle f, g\rangle = 0$. f and g are orthogonal if and only if $\|f+g\|_2^2 = \|f\|_2^2 + \|g\|_2^2$. For centred (mean zero) random variables X, Y, we have $\langle X, Y\rangle = \mathbb{E}[XY] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathrm{Cov}(X, Y)$ which vanishes when X and Y are orthogonal.

Example. The parallelogram identity holds: $\|f+g\|_2^2 + \|f-g\|_2^2 = 2\big(\|f\|_2^2 + \|g\|_2^2\big)$

Definition. Let $V \subseteq L^2(\mu)$. We define its *orthogonal complement* to be

$$V^\perp = \big\{f \in L^2(\mu) \mid \forall g \in V, \langle f, g\rangle = 0\big\}$$

We say that a subset V of $\mathcal{L}^2$ is *closed* if any sequence $f_n \in V$ that converges in $\mathcal{L}^2$, its limit f coincides almost everywhere with some $v \in V$.

Theorem. Let V be a closed linear subspace of $\mathcal{L}^2(\mu)$. Then for all $f \in \mathcal{L}^2$, there exists an orthogonal decomposition $f = v+u$ where $v \in V$ and $u \in V^\perp$ such that $\|f - v\|_2 \leq \|f - g\|_2$ for all $g \in V$ with equality only if $v = g$ almost everywhere. We call v the *projection* of f onto V.

Proof. In this proof, we set $p = 2$ for all norms. We define $d(f, V) = \inf_{g \in V} \|g - f\|$, and let $g_n \in V$ be a sequence of functions such that $\|g_n - f\|$ converges to $d(f, V)$. By the parallelogram law,

$$\begin{aligned} 2\|f - g_n\|^2 + 2\|f - g_m\|^2 &= \|2f - (g_n + g_m)\|^2 + \|g_n - g_m\|^2 \\ &= 4\left\|f - \underbrace{\frac{g_n + g_m}{2}}_{\in V}\right\|^2 + \|g_n - g_m\|^2 \\ &\geq 4d(f, V)^2 + \|g_n - g_m\|^2 \end{aligned}$$

Taking the limit superior as $n, m \to \infty$, $\limsup_{m,n} \|g_n - g_m\|^2 \leq 4d(f, V) - 4d(f, V) = 0$. So the sequence g_n is Cauchy in L^2, so by completeness, it converges to some $v \in L^2$. Since V is closed, $v \in V$. In particular, $d(f, V) = \inf_{g \in V} \|g - f\| = \|v - f\|$.

Note that $d(f, V)^2 \leq F(t) = \|f - (v + th)\|^2$ where $t \in \mathbb{R}$ and $h \in V$. So we obtain the first-order condition $F'(0) = 2\langle f - v, h\rangle = 0$ for all h. Defining $f - v = u$, we have $f = u+v$ and $u \in V^\perp$ since h was arbitrary.

For uniqueness, suppose $f = w + z$ with $w \in V$ and $z \in V^\perp$. Then $v - w + u - z = f - f = 0$, so taking norms, $0 = \|v - w + u - z\|^2 = \|v - w\|^2 + \|u - z\|^2$ so $v = w$ and $u = z$ (almost everywhere) by orthogonality. □

5.4. Convergence in probability and uniform integrability

Theorem (bounded convergence). Let X_n be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $|X_n| \leq C < \infty$ and they converge in probability to X. Then $X_n \to X$ in $L^1(\mathbb{P})$.

Proof. We know that $X_{n_k} \to X$ almost surely along a subsequence n_k. So $|X| = \lim_k |X_{n_k}| \leq C < \infty$ almost surely. Then

$$\begin{aligned} \mathbb{E}\left[|X_n - X|\right] &= \mathbb{E}\left[|X_n - X|\left(\mathbb{1}_{\{|X_n - X| > \frac{\varepsilon}{2}\}} + \mathbb{1}_{\{|X_n - x| \leq \frac{\varepsilon}{2}\}}\right)\right] \\ &\leq 2C\mathbb{P}\left(|X_n - X| \geq \frac{\varepsilon}{2}\right) + \frac{\varepsilon}{2} \\ &< \varepsilon \end{aligned}$$

for sufficiently large n. □

If $X \in L^1(\mathbb{P})$, then as $\delta \to 0$,

$$I_X(\delta) = \sup\{\mathbb{E}\left[|X|\mathbb{1}_A\right] \mid \mathbb{P}(A) \leq \delta\} \downarrow 0$$

Suppose this does not hold. Then there exists $\varepsilon > 0$ and a sequence of events $A_n \in \mathcal{F}$ such that $\mathbb{P}(A_n) \leq 2^{-n}$ but $\mathbb{E}\left[|X|\mathbb{1}_{A_n}\right] \geq \varepsilon$. Since $\sum_n \mathbb{P}(A_n) < \infty$, by the first Borel–Cantelli lemma, we have $\mathbb{P}\left(\bigcap_n \bigcup_{m \geq n} A_m\right) = 0$. But $\mathbb{E}\left[|X|\mathbb{1}_{A_n}\right] \leq \mathbb{E}\left[|X|\mathbb{1}_{\bigcup_{m \geq n} A_m}\right]$. Note that $\mathbb{1}_{\bigcup_{m \geq n} A_m} \to \mathbb{1}_{\bigcap_n \bigcup_{m \geq n} A_n}$, so $\mathbb{E}\left[|X|\mathbb{1}_{\bigcup_{m \geq n} A_m}\right] \to \mathbb{E}\left[|X|\mathbb{1}_{\bigcap_n \bigcup_{m \geq n}}\right]$ by the dominated convergence theorem, but this is equal to zero, giving a contradiction.

Definition. For a collection $\mathcal{X} \subseteq L^1(\mathbb{P})$ of random variables, we say $\mathcal{X}$ is *uniformly integrable* if it is bounded in $L^1(\mathbb{P})$, and

$$I_{\mathcal{X}}(\delta) = \sup\{\mathbb{E}\left[|X|\mathbb{1}_A\right] \mid \mathbb{P}(A) \leq \delta, X \in \mathcal{X}\} \downarrow 0$$

Remark. Note that $X_n = n\mathbb{1}_{\left[0, \frac{1}{n}\right]}$ for the Lebesgue measure μ on $[0, 1]$ is bounded in $L^1(\mathbb{P})$ but not uniformly integrable. If $\mathcal{X}$ is bounded in $L^p(\mathbb{P})$ for $p > 1$, then by Hölder's inequality,

$$\mathbb{E}\left[|X|\mathbb{1}_A\right] \leq \underbrace{\|X\|_p}_{\text{bounded}} \cdot \underbrace{\mathbb{P}(A)^{\frac{1}{q}}}_{\leq \delta^{\frac{1}{q}} \to 0}$$

Lemma. $\mathcal{X} \subseteq L^1(\mathbb{P})$ is uniformly integrable if and only if $\sup_{X \in \mathcal{X}} \mathbb{E}\left[|X|\mathbb{1}_{\{|X| > K\}}\right] \to 0$ as $K \to \infty$.

Proof. Let $\mathcal{X}$ be uniformly integrable. Applying Markov's inequality, as $K \to \infty$,

$$\mathbb{P}(|X| > K) \leq \frac{\mathbb{E}[|X|]}{K} = \frac{\mathbb{E}[|X|\mathbb{1}_\Omega]}{K} \leq \frac{I_{\mathcal{X}}(1)}{K} \to 0$$

Using the uniform integrability property using $A = \{|X| > K\}$, we obtain the required limit. Conversely, we have

$$\mathbb{E}[|X|] = \mathbb{E}\left[|X|\left(\mathbb{1}_{\{|X|\leq K\}} + \mathbb{1}_{\{|X|>K\}}\right)\right] \leq K + \frac{\varepsilon}{2}$$

for sufficiently large K. So $\mathcal{X}$ is bounded in $L^1(\mathbb{P})$ as required. Then for A such that $\mathbb{P}(A) \leq \delta$,

$$\mathbb{E}\left[|X|\mathbb{1}_A\left(\mathbb{1}_{\{|X|\leq K\}} + \mathbb{1}_{\{|X|>K\}}\right)\right] \leq K\mathbb{P}(A) + \mathbb{E}\left[|X|\mathbb{1}_{\{|X|>K\}}\right] \leq K\delta + \frac{\varepsilon}{2} < \varepsilon$$

for sufficiently small δ. □

Theorem. Let X_n, X be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Then the following are equivalent.

(i) $X_n, X \in L^1(\mathbb{P})$ and $X_n \to X$ in $L^1(\mathbb{P})$.

(ii) $\{X_n \mid n \in \mathbb{N}\}$ is uniformly integrable, and $X_n \to X$ in probability.

Proof. (i) implies (ii). Using Markov's inequality,

$$\mathbb{P}(|X_n - X| > \varepsilon) \leq \frac{\mathbb{E}[|X_n - X|]}{\varepsilon} \to 0$$

so $X_n \to X$ in probability. Since any finite collection is uniformly integrable, so are X along with $X_1, \dots, X_N$ for each N. For the indices larger than N, we have

$$\mathbb{E}[|X_n|\mathbb{1}_A] \leq \mathbb{E}[|X_n - X|\mathbb{1}_A] + \mathbb{E}[|X|\mathbb{1}_A] \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2}$$

for sufficiently large N and sufficiently small δ, so all X_n are uniformly integrable.

(ii) implies (i). Along a subsequence, $X_n \to X$ almost surely. So

$$\mathbb{E}[|X|] = \mathbb{E}\left[\liminf_k |X_{n_k}|\right] \leq \liminf_k \mathbb{E}\left[|X_{n_k}|\right] \leq I_{\mathcal{X}}(1) < \infty$$

almost surely, so $X \in L^1(\mathbb{P})$. Next, we define random variables $g(X_n) = X_n^K = \max(-K, \min(K, X_n))$ and $g(X) = X^K = \max(-K, \min(K, X))$, where g is continuous. Then for some $\varepsilon' > 0$,

$$\mathbb{P}(|g(X_n) - g(X)| > \varepsilon) \leq \mathbb{P}(|X_n - X| > \varepsilon') \to 0$$

as $n \to \infty$, since $X_n \to X$ in probability and g is continuous. Then by bounded convergence, $X_n^K \to X^K$ in L^1, and so

$$\begin{aligned}
\mathbb{E}[|X_n - X|] &\leq \mathbb{E}\left[|X_n - X_n^K|\right] + \mathbb{E}\left[|X_n^K - X^K|\right] + \mathbb{E}\left[|X^K - X|\right] \\
&= \mathbb{E}\left[|X_n|\mathbb{1}_{\{|X_n|>k\}}\right] + \mathbb{E}\left[|X_n^K - X^K|\right] + \mathbb{E}\left[|X|\mathbb{1}_{\{|X|>K\}}\right] \\
&< \varepsilon
\end{aligned}$$

by choosing sufficiently large K and n. □

6. Fourier analysis

6.1. Fourier transforms

In this section, we will write $L^p(\mathbb{R}^d)$ for the set of measurable functions $f : \mathbb{R}^d \to \mathbb{C}$ such that $\|f\|_p = \left(\int_{\mathbb{R}^d} |f(x)|^p \,\mathrm{d}x\right)^{\frac{1}{p}} < \infty$. We can extend the integral as a complex linear map $L^1(\mathbb{R}) \to \mathbb{C}$ by defining

$$\int_{\mathbb{R}} (u + iv)(x)\,\mathrm{d}x = \int_{\mathbb{R}} u(x)\,\mathrm{d}x + i\int_{\mathbb{R}} v(x)\,\mathrm{d}x$$

Note that for some $u + iv = \alpha \in \mathbb{C}$ with $|\alpha| = 1$,

$$\left|\int_{\mathbb{R}^d} f(x)\,\mathrm{d}x\right| = \int_{\mathbb{R}^d} \alpha f(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} u(x)\,\mathrm{d}x + i\int_{\mathbb{R}^d} v(x)\,\mathrm{d}x$$

But since the left hand side is real-valued, the $i\int_{\mathbb{R}^d} v(x)\,\mathrm{d}x$ term vanishes. So

$$\left|\int_{\mathbb{R}^d} f(x)\,\mathrm{d}x\right| = \int_{\mathbb{R}^d} u(x)\,\mathrm{d}x \leq \int_{\mathbb{R}^d} |f(x)|\,\mathrm{d}x$$

Definition. Let $f \in L^1(\mathbb{R}^d)$. We define the *Fourier transform* $\hat{f}$ by

$$\hat{f}(u) = \int_{\mathbb{R}^d} f(x)e^{i\langle u,x\rangle}\,\mathrm{d}x$$

where $\langle u, x\rangle = \sum_{i=1}^d u_i x_i$.

Remark. Note that $\left|\hat{f}(u)\right| \leq \|f\|_1$. Also, if $u_n \to u$, then $e^{i\langle u_n,x\rangle} \to e^{i\langle u,x\rangle}$. By the dominated convergence theorem with dominating function $|f|$, we have $\hat{f}(u_n) \to \hat{f}(u)$, so $\hat{f}$ is a continuous bounded function.

Definition. Let $f \in L^1(\mathbb{R}^d)$ such that $\hat{f} \in L^1(\mathbb{R}^d)$. Then we say that the *Fourier inversion formula* holds for f if

$$f(x) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \hat{f}(u)e^{-i\langle u,x\rangle}\,\mathrm{d}u$$

almost everywhere in $\mathbb{R}^d$.

Definition. Let $f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$. Then the *Plancherel identity* holds for f if

$$\left\|\hat{f}\right\|_2 = (2\pi)^{\frac{d}{2}}\|f\|_2$$

We will show that the Fourier inversion formula holds whenever $\hat{f} \in L^1(\mathbb{R}^d)$, and the Plancherel identity holds for all $f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$.

Remark. Given the Plancherel identity, the Fourier transform is a linear isometry of $L^2(\mathbb{R}^d)$, by approximating any function in $L^2(\mathbb{R}^d)$ by integrable functions.

Definition. Let μ be a finite Borel measure on $\mathbb{R}^d$. We define the Fourier transform of the measure by

$$\hat{\mu}(u) = \int_{\mathbb{R}^d} e^{i\langle u,x\rangle}\,\mathrm{d}\mu(x)$$

Note that $|\hat{\mu}(u)| \leq \mu(\mathbb{R}^d)$, and $\hat{\mu}$ is continuous by the dominated convergence theorem. If μ has a density f with respect to the Lebesgue measure, $\hat{\mu} = \hat{f}$.

Definition. Let X be an $\mathbb{R}^d$-valued random variable. The *characteristic function* φ_X is given by

$$\varphi_X(u) = \mathbb{E}\left[e^{i\langle u,X\rangle}\right] = \hat{\mu}_X(u)$$

where μ_X is the law of X.

6.2. Convolutions

Definition. Let $f \in L^1(\mathbb{R}^d)$ and ν be a probability measure on $\mathbb{R}^d$. We define their *convolution $f * \nu$* by

$$(f * \nu)(x) = \begin{cases} \int_{\mathbb{R}^d} f(x-y)\,\mathrm{d}\nu(y) & \text{if } (y \mapsto f(x-y)) \in L^1(\nu) \\ 0 & \text{else} \end{cases}$$

Remark. If $1 \leq p < \infty$, by Jensen's inequality,

$$\begin{aligned} \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} |f(x-y)|\,\mathrm{d}\nu(y) \right)^p \mathrm{d}x &\leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |f(x-y)|^p\,\mathrm{d}\nu(y)\,\mathrm{d}x \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |f(x-y)|^p\,\mathrm{d}x\,\mathrm{d}\nu(y) \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |f(x)|\,\mathrm{d}\nu(y)\,\mathrm{d}x \\ &= \int_{\mathbb{R}^d} |f(x)|\,\mathrm{d}x \\ &= \|f\|_p^p \end{aligned}$$

So $f \in L^p(\mathbb{R}^d)$, we have $(y \mapsto f(x-y)) \in L^p(\nu)$ almost everywhere, and again by Jensen's inequality,

$$\|f * \nu\|_p^p = \int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} f(x-y)\,\mathrm{d}\nu(y) \right|^p \mathrm{d}x \leq \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} |f(x-y)|\,\mathrm{d}\nu(y) \right)^p \mathrm{d}x \leq \|f\|_p^p$$

Hence $f \mapsto f * \nu$ is a contraction on $L^p(\mathbb{R}^d)$.

In the case where ν has a density g with respect to the Lebesgue measure, we write $f * g = f * \nu$.

Definition. For probability measures μ, ν on $\mathbb{R}^d$, their convolution $\mu * \nu$ is a probability measure on $\mathbb{R}^d$ given by the law of $X + Y$ where X, Y are independent random variables with laws μ and ν, so

$$\begin{aligned}(\mu * \nu)(A) &= \mathbb{P}(X + Y \in A) \\ &= \int_{\mathbb{R}^d \times \mathbb{R}^d} \mathbb{1}_A(x+y) \,\mathrm{d}(\mu \otimes \nu)(x,y) \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \mathbb{1}_A(x+y) \,\mathrm{d}\nu(y) \,\mathrm{d}\mu(x)\end{aligned}$$

If μ has density f with respect to the Lebesgue measure, $\mu * \nu$ has density $f * \nu$ with respect to the Lebesgue measure. Indeed,

$$\begin{aligned}(\mu * \nu)(A) &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \mathbb{1}_A(x+y) f(x) \,\mathrm{d}x \,\mathrm{d}\nu(y) \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \mathbb{1}_A(v) f(v-y) \,\mathrm{d}v \,\mathrm{d}\nu(y) \\ &= \int_{\mathbb{R}^d} \mathbb{1}_A(v) \int_{\mathbb{R}^d} f(v-y) \,\mathrm{d}\nu(y) \,\mathrm{d}v \\ &= \int_{\mathbb{R}^d} \mathbb{1}_A(v) (f * \nu)(v) \,\mathrm{d}v\end{aligned}$$

Proposition. $\widehat{f * \nu}(u) = \hat{f}(u)\hat{\nu}(u)$.

Proposition. $\widehat{\mu * \nu}(u) = \mathbb{E}\left[e^{i\langle u, X+Y\rangle}\right] = \mathbb{E}\left[e^{i\langle u,X\rangle} e^{i\langle u,Y\rangle}\right] = \hat{\mu}(u)\hat{\nu}(u)$.

6.3. Fourier transforms of Gaussians

Definition. The *normal distribution* $N(0,t)$ is given by the probability density function

$$g_t(x) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}$$

If φ_X is the characteristic function of a standard normal random variable, by integration by parts,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}u}\varphi_X(u) &= \frac{\mathrm{d}}{\mathrm{d}u}\int_{\mathbb{R}} e^{iux} g_1(x)\,\mathrm{d}x \\ &= \int_{\mathbb{R}} g_1(x)\frac{\mathrm{d}}{\mathrm{d}u}e^{iux}\,\mathrm{d}x \\ &= \frac{i}{\sqrt{2\pi}}\int_{\mathbb{R}} \underbrace{e^{iux}}_{v}\underbrace{xe^{-\frac{x^2}{2}}}_{w'}\,\mathrm{d}x \\ &= \frac{i^2}{\sqrt{2\pi}}\int_{\mathbb{R}} ue^{iux}e^{-\frac{x^2}{2}}\,\mathrm{d}x \\ &= -u\varphi_X(u)\end{aligned}$$

Hence,

$$\frac{\mathrm{d}}{\mathrm{d}u}\left(e^{\frac{u^2}{2}}\varphi_X(u)\right) = ue^{\frac{u^2}{2}}\varphi_X(u) - e^{\frac{u^2}{2}}u\varphi_X(u) = 0$$

In particular, $\varphi_X(u) = \varphi_X(0)e^{-\frac{u^2}{2}} = e^{-\frac{u^2}{2}}$. In other words, $\hat{g}_1(u) = \sqrt{2\pi}g_1(u)$.

In $\mathbb{R}^d$, consider a Gaussian random vector $Z = (Z_1, \dots, Z_d)$ with independent and identically distributed entries $Z_i \sim N(0,1)$. Then, the joint probability density function of $\sqrt{t}Z$ is

$$g_t(x) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi t}}e^{-\frac{x_j^2}{2t}} = (2\pi t)^{-\frac{d}{2}}e^{-\frac{\|x\|^2}{2t}}$$

The Fourier transform of g_t is

$$\hat{g}_t(u) = \mathbb{E}\left[e^{i\langle u,\sqrt{t}Z\rangle}\right] = \mathbb{E}\left[\prod_{j=1}^{d} e^{iu_j\sqrt{t}z_j}\right] = \prod_{j=1}^{d}\mathbb{E}\left[e^{iu_j\sqrt{t}z_j}\right] = \prod_{j=1}^{d} e^{-u_j^2\frac{t}{2}} = e^{-\frac{\|u\|^2 t}{2}}$$

which implies that in general, $\hat{g}_t(u) = (2\pi)^{\frac{d}{2}}t^{\frac{d}{2}}g_{\frac{1}{t}}(u)$. Taking the Fourier transform with respect to u, $\hat{\hat{g}}_t = (2\pi)^d g_t$, and since $g_t(-x) = g_t(x)$ and the Lebesgue measure is translation invariant, we have

$$g_t(x) = \frac{1}{(2\pi)^d}\hat{\hat{g}}_t(x) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{-i\langle u,x\rangle}\hat{g}_t(u)\,\mathrm{d}u$$

so the Fourier inversion theorem holds for such Gaussian random vectors.

Definition. We say that a function on $\mathbb{R}^d$ is a *Gaussian convolution* if it is of the form

$$f * g_t(x) = \int_{\mathbb{R}^d} f(x-y)g_t(y)\,\mathrm{d}y$$

where $x \in \mathbb{R}^d, t > 0, f \in L^1(\mathbb{R}^d)$.

We can show that $f * g_t$ is continuous on $\mathbb{R}^d$, and $\|f * g_t\|_1 \leq \|f\|_1$. Note that $\widehat{f * g_t}(u) = \hat{f}(u)e^{-\frac{\|u\|^2 t}{2}}$, so $\left\|\widehat{f * g_t}\right\|_\infty \leq \|f\|_1$, giving $\left\|\widehat{f * g_t}\right\|_1 \leq \|f\|_1 (2\pi)^{\frac{d}{2}} t^{-\frac{d}{2}} < \infty$.

Lemma. The Fourier inversion theorem holds for all Gaussian convolutions.

Proof. We can use the Fourier inversion theorem for $g_t(y)$ to see that

$$\begin{aligned}
(2\pi)^d f * g_t(x) &= (2\pi)^d \int_{\mathbb{R}^d} f(x-y) g_t(y)\,\mathrm{d}y \\
&= \int_{\mathbb{R}^d} f(x-y) \int_{\mathbb{R}^d} e^{-i\langle u,y\rangle} \hat{g}_t(u)\,\mathrm{d}u\,\mathrm{d}y \\
&= \int_{\mathbb{R}^d} e^{-i\langle u,x\rangle} \int_{\mathbb{R}^d} f(x-y) e^{i\langle u,x-y\rangle}\,\mathrm{d}y\,\hat{g}_t(u)\,\mathrm{d}u \\
&= \int_{\mathbb{R}^d} e^{-i\langle u,x\rangle} \int_{\mathbb{R}^d} f(z) e^{i\langle u,z\rangle}\,\mathrm{d}z\,\hat{g}_t(u)\,\mathrm{d}u \\
&= \int_{\mathbb{R}^d} e^{-i\langle u,x\rangle} \hat{f}(u)\hat{g}_t(u)\,\mathrm{d}u \\
&= \int_{\mathbb{R}^d} e^{-i\langle u,x\rangle} \widehat{f * g_t}(u)\,\mathrm{d}u
\end{aligned}$$

□

Remark. If μ is a finite measure, then $\mu * g_t = \mu * g_{\frac{t}{2}} * g_{\frac{t}{2}}$ with $\mu * g_{\frac{t}{2}} \in L^1$, so is also a Gaussian convolution.

Lemma (Gaussian convolutions are dense in L^p)**.** Let $f \in L^p$ where $1 \leq p < \infty$. Then $\|f * g_t - f\|_p \to 0$ as $t \to 0$.

Proof. One can easily show that the space $C_c(\mathbb{R}^d)$ of continuous functions of compact support is dense in L^p. Hence, for all $\varepsilon > 0$, there exists $h \in C_c(\mathbb{R}^d)$ such that $\|f - h\|_p < \frac{\varepsilon}{3}$, and by properties of the convolution, we also obtain

$$\|f * g_t - h * g_t\|_p = \|(f-h) * g_t\|_p \leq \|f - h\|_p < \frac{\varepsilon}{3}$$

So

$$\|f * g_t - f\|_p \leq \|f * g_t - h * g_t\|_p + \|h * g_t + h\|_p + \|h - f\|_p < \frac{\varepsilon}{2} + \|h * g_t - h\|_p$$

so it suffices to prove the result for $f = h \in C_c(\mathbb{R}^d)$. We define a new map

$$e(y) = \int_{\mathbb{R}^d} |h(x-y) - h(x)|^p\,\mathrm{d}x$$

Since h is bounded on its bounded support, the dominated convergence theorem implies that e is continuous at $y = 0$. Note that $e(y) \leq 2^{p+1}\|h\|_p^p$. Hence, by Jensen's inequality,

$$\begin{aligned}\|h * g_t - h\|_p^p &= \int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} (h(x-y) - h(x)) g_t(y) \,\mathrm{d}y \right|^p \mathrm{d}x \\ &\leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |h(x-y) - h(x)|^p \,\mathrm{d}x\, g_t(y) \,\mathrm{d}y \\ &= \int_{\mathbb{R}^d} e(y) g_t(y) \,\mathrm{d}y \\ &= \int_{\mathbb{R}^d} \underbrace{e(\sqrt{t}z)}_{\to e(0)=0 \text{ as } t \to 0} g_1(z) \,\mathrm{d}z \\ &\to 0\end{aligned}$$

□

Theorem (Fourier inversion)**.** Let $f \in L^1(\mathbb{R}^d)$ be such that $\hat{f} \in L^1(\mathbb{R}^d)$. Then for almost all $x \in \mathbb{R}^d$,

$$f(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i\langle u,x\rangle} \hat{f}(u) \,\mathrm{d}u$$

Remark. This proves that the Fourier transform is injective; $\hat{f} = \hat{g}$ implies $\widehat{f - g} = 0$ so by Fourier inversion, $f = g$ almost everywhere. The identity holds everywhere on $\mathbb{R}^d$ for the (unique) continuous representative f in its equivalence class.

Proof. The Fourier inversion theorem holds for the following Gaussian convolution for all t.

$$f * g_t(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i\langle u,x\rangle} \hat{f}(u) e^{\frac{-|u|^2 t}{2}} \,\mathrm{d}u = f_t(x)$$

Now, since Gaussian convolutions are dense, $f * g_t \to f$ in L^1, so $f * g_t \to f$ in measure by Markov's inequality. Hence, along a subsequence, $f * g_{t_k} \to f$ almost everywhere. On the other hand, by the dominated convergence theorem with dominating function $|\hat{f}|$, the right hand side converges to $\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i\langle u,x\rangle} \hat{f}(u) \,\mathrm{d}u$. So this is equal to $\lim_{t_k \to 0} f_{t_k}$ almost everywhere by uniqueness of limits. □

Theorem (Plancherel)**.** Let $f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$. Then $\|f\|_2 = (2\pi)^{-\frac{d}{2}} \left\|\hat{f}\right\|_2$.

Remark. By the Pythagorean identity, $\langle f, g\rangle = (2\pi)^{-d} \langle \hat{f}, \hat{g}\rangle$.

Proof. Initially, we assume $\hat{f} \in L^1$. In this case, $f, \hat{f} \in L^\infty$, and $(x, u) \mapsto f(x)\hat{f}(u)$ is integrable for the product Lebesgue measure $\mathrm{d}x \otimes \mathrm{d}u$ on $\mathbb{R}^d \times \mathbb{R}^d$, so Fubini's theorem for bounded functions applies.

$$\begin{aligned}(2\pi)^d\|f\|_2^2 &= (2\pi)^d \int_{\mathbb{R}^d} f(x)\overline{f(x)}\,\mathrm{d}x \\ &= \int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} e^{-i\langle u,x\rangle}\hat{f}(u)\,\mathrm{d}u\right)\overline{f(x)}\,\mathrm{d}x \\ &= \int_{\mathbb{R}^d}\hat{f}(u)\overline{\int_{\mathbb{R}^d} e^{i\langle u,x\rangle}f(x)\,\mathrm{d}x}\,\mathrm{d}u \\ &= \int_{\mathbb{R}^d}\hat{f}(u)\overline{\hat{f}(u)}\,\mathrm{d}u \\ &= \left\|\hat{f}\right\|_2^2\end{aligned}$$

To extend this result to general f, we take the Gaussian convolutions $f * g_t = f_t$ such that $f_t \to f$ in L^2. By the continuity of the norm, $\|f_t\|_2 \to \|f\|_2$. Since $\left|\hat{f}(u)e^{-\frac{|u|^2 t}{2}}\right|^2$ increases to $\left|\hat{f}(u)\right|^2$, we have by monotone convergence that $\left\|\hat{f}_t\right\|_2^2 \uparrow \left\|\hat{f}\right\|_2^2$. Therefore, since the Plancherel identity holds for the f_t,

$$\|f\|_2^2 = \lim_{t\to 0}\|f_t\|_2^2 = \lim_{t\to 0}(2\pi)^{-d}\left\|\hat{f}_t\right\|_2^2 = (2\pi)^{-d}\left\|\hat{f}\right\|_2^2$$

□

Remark. Since $L_1 \cap L_2$ is dense in L^2, we can extend the linear operator $F_0(f) = (2\pi)^{-\frac{d}{2}}\hat{f}$ to L^2 by continuity to a linear isometry $F : L^2 \to L^2$ known as the *Fourier–Plancherel transform*. One can show that F is surjective with inverse $F^{-1} : L^2 \to L^2$.

Example. Consider the Dirac measure δ_0 on $\mathbb{R}$, so $\hat{\delta}_0(u) = \int_{\mathbb{R}} e^{iux}\,\mathrm{d}\delta_0(x) = 1$. But the inverse Fourier transform would be $\frac{1}{2\pi}\int_{\mathbb{R}} e^{iux}\,\mathrm{d}u$ which is not a Lebesgue integrable function.

Theorem. Let X be a random vector in $\mathbb{R}^d$ with law μ_X. Then the characteristic function $\varphi_X = \hat{\mu}_X$ uniquely determines μ_X. In addition, if $\varphi_X \in L^1$, then μ_X has a probability density function f_X which can be computed almost everywhere by $\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{-i\langle u,x\rangle}\varphi_X(u)\,\mathrm{d}u$.

Proof. Let $Z = (Z_1, \dots, Z_d)$ be a vector of independent and identically distributed random variables, independent of X, with $Z_j \sim N(0,1)$. Then $\sqrt{t}Z$ has probability density function g_t. Then $X + \sqrt{t}Z$ has probability density function $f_t = \mu_X * g_t$. This is a Gaussian convolution since $\mu_X * g_t = \mu_X * g_{\frac{t}{2}} * g_{\frac{t}{2}}$. Hence,

$$f_t(x) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{i\langle u,x\rangle}\varphi_X(u)e^{-\frac{|u|^2 t}{2}}\,\mathrm{d}u$$

which is uniquely determined by φ_X. We show on an example sheet that two Borel probability measures μ, ν on $\mathbb{R}^d$ coincide if and only if $\mu(g) = \nu(g)$ for all $g : \mathbb{R}^d \to \mathbb{R}$ that are bounded, continuous, and have compact support. Now,

$$\int_{\mathbb{R}^d} g(x) f_t(x)\,\mathrm{d}x = \mathbb{E}\left[\underbrace{g(X + \sqrt{t}Z)}_{\to X \text{ a.s.}}\right]$$

Since $\left|g(X + \sqrt{t}Z)\right| \leq \|g\|_\infty < \infty$, by the bounded convergence theorem, this converges to $\mathbb{E}\left[g(X)\right] = \int_{\mathbb{R}^d} g(x)\,\mathrm{d}\mu_X(x)$. So by uniqueness of limits, φ_X determines μ_X.

If $\varphi_X \in L^1$, by dominated convergence, $f_t(x)$ converges everywhere to some function f_X. In particular, since $\mu_X * g_t \geq 0$, the limit f_X is also nonnegative on $\mathbb{R}^d$. Then, for any bounded continuous function on compact support $g \in C_c^b(\mathbb{R}^d)$,

$$\int_{\mathbb{R}^d} g(x) f_X(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} g(x) \lim_{t \to 0} \underbrace{f_t(x)}_{\|\varphi_X\|_1}\,\mathrm{d}x = \lim_{t \to 0} \int_{\mathbb{R}^d} g(x) f_t(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} g(x)\,\mathrm{d}\mu_X(x)$$

by the dominated convergence theorem, since g has compact support. □

Definition. A sequence $(\mu_n)_{n \in \mathbb{N}}$ of Borel probability measures on $\mathbb{R}^d$ *converges weakly* to a Borel probability measure μ if $\mu_n(g) \to \mu(g)$ for all $g : \mathbb{R}^d \to \mathbb{R}$ bounded and continuous. If $(X_n)_{n \in \mathbb{N}}, X$ are random vectors with laws $(\mu_{X_n}), \mu_X$ such that μ_{X_n} converges weakly to μ_X, we say (X_n) converges weakly to X.

Remark. If $d = 1$, weak convergence is equivalent to convergence in distribution; this is proven on an example sheet. One can also show that convergence of $\mu_n(g)$ to $\mu(g)$ for all $g \in C_c^\infty(\mathbb{R}^d)$ suffices to show weak convergence, where $C_c^\infty(\mathbb{R}^d)$ is the space of smooth functions of compact support. This is equivalent to the notion of weak-$\star$ convergence on the function space $C_b(\mathbb{R}^d)$.

Theorem (Lévy's continuity theorem). Let X_n, X be random vectors in $\mathbb{R}^d$, such that $\varphi_{X_n}(u) \to \varphi_X(u)$ for all u, as $n \to \infty$. Then $\mu_{X_n} \to \mu_X$ weakly.

Remark. The converse holds by definition of weak convergence, testing against the complex exponentials in the Fourier transform.

Proof. Let $Z = (Z_1, \dots, Z_d)$ be a vector of standard normal random variables, independent from each other, X_n, and X. Let $g \in C_c^\infty(\mathbb{R}^d)$. Then $g \in L^1(\mathbb{R}^d)$, and is Lipschitz by the mean value theorem, as its first derivative is bounded. Let $|g(x) - g(y)| \leq \|g\|_{\mathrm{Lip}}|x - y|$. Let $\varepsilon > 0$.

Let $t > 0$ be sufficiently small such that $\sqrt{t}\|g\|_{\mathrm{Lip}}\mathbb{E}\left[|Z|\right] < \frac{\varepsilon}{3}$. Then,

$$\begin{aligned}
\left|\mu_{X_n}(g) - \mu_X(g)\right| &= \left|\mathbb{E}\left[g(X_n)\right] - \mathbb{E}\left[g(X)\right]\right| \\
&\leq \mathbb{E}\left[\left|g(X_n) - g(X_n + \sqrt{t}Z)\right|\right] + \mathbb{E}\left[\left|g(X) - g(X + \sqrt{t}Z)\right|\right] \\
&+ \left|\mathbb{E}\left[g(X_n + \sqrt{t}Z) - g(X + \sqrt{t}Z)\right]\right| \\
&\leq 2\|g\|_{\mathrm{Lip}}\sqrt{t}\mathbb{E}\left[|Z|\right] + \left|\mathbb{E}\left[g(X_n + \sqrt{t}Z) - g(X + \sqrt{t}Z)\right]\right| \\
&\leq \frac{2\varepsilon}{3} + \left|\mathbb{E}\left[g(X_n + \sqrt{t}Z) - g(X + \sqrt{t}Z)\right]\right|
\end{aligned}$$

We show that the remaining term can be made less than $\frac{\varepsilon}{3}$ as $n \to \infty$. Let $f_{t,n}(x) = g_t * \mu_{X_n}$. Then, by Fourier inversion for Gaussian convolutions,

$$\begin{aligned}
\mathbb{E}\left[g(X_n + \sqrt{t}Z)\right] &= \int_{\mathbb{R}^d} g(x) f_{t,n}(x)\,\mathrm{d}x \\
&= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} g(x) \int_{\mathbb{R}^d} e^{-i\langle u,x\rangle}\varphi_{X_n}(u) e^{-\frac{|u|^2 t}{2}}\,\mathrm{d}u\,\mathrm{d}x
\end{aligned}$$

Since characteristic functions are bounded by 1, we can apply the dominated convergence theorem with dominating function $|g(x)|e^{-\frac{|u|^2 t}{2}}$ to find

$$\begin{aligned}
\mathbb{E}\left[g(X_n + \sqrt{t}Z)\right] &\to \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} g(x) \int_{\mathbb{R}^d} e^{-i\langle u,x\rangle}\varphi_{X}(u) e^{-\frac{|u|^2 t}{2}}\,\mathrm{d}u\,\mathrm{d}x \\
&= \int_{\mathbb{R}^d} g(x) f_t(x)\,\mathrm{d}x \\
&= \mathbb{E}\left[g(X + \sqrt{t}Z)\right]
\end{aligned}$$

where $f_t = g_t * \mu_X$. So as $n \to \infty$, the difference between these two terms can be made less than $\frac{\varepsilon}{3}$ as required. □

Theorem (central limit theorem). Let $X_1, \dots, X_n$ be independent and identically distributed random variables with $\mathbb{E}\left[X_i\right] = 0$ and $\mathrm{Var}\left(X_i\right) = 1$. Let $S_n = \sum_{i=1}^n X_n$. Then

$$\frac{1}{\sqrt{n}} S_n \xrightarrow{\text{weakly}} Z \sim N(0,1)$$

In particular,

$$\mathbb{P}\left(\frac{1}{\sqrt{n}} S_n \leq x\right) \to \mathbb{P}\left(Z \leq x\right)$$

Proof. Let $X = X_1$. The characteristic function $\varphi(u) = \varphi_X(u) = \mathbb{E}\left[e^{iuX}\right]$ satisfies $\varphi(0) = 1$, $\varphi'(u) = i\mathbb{E}\left[Xe^{iuX}\right]$, $\varphi''(u) = i^2\mathbb{E}\left[X^2 e^{iuX}\right]$. We can find $\varphi'(0) = i\mathbb{E}\left[X\right] = 0$ and $\varphi''(0) =$

$-\mathbb{E}\left[X^2\right] = -\operatorname{Var}(X) = -1$. By Taylor's theorem, $\varphi(v) = 1 - \frac{v^2}{2} + o(v^2)$ as $v \to 0$. Now, denoting $\varphi_n(u) = \varphi_{\frac{1}{\sqrt{n}}S_n}(u)$, we can write

$$\begin{aligned}
\varphi_n(u) &= \mathbb{E}\left[e^{i\frac{u}{\sqrt{n}}(X_1+\dots+X_n)}\right] \\
&= \prod_{j=1}^{n} \mathbb{E}\left[e^{i\frac{u}{\sqrt{n}}X_j}\right] \\
&= \left[\varphi\left(\frac{u}{\sqrt{n}}\right)\right]^n \\
&= \left[1 - \frac{u^2}{2n} + o\left(\frac{1}{n}\right)\right]^n
\end{aligned}$$

The complex logarithm satisfies $\log(1+z) = z + o(z)$, so by taking logarithms, we find

$$\log \varphi_n(u) = n \log\left(1 - \frac{u^2}{2n} + o\left(\frac{1}{n}\right)\right) = -\frac{u^2}{2}$$

Hence, $\varphi_n(u) \to e^{-\frac{|u|^2}{2}} = \varphi_Z(u)$. So by Lévy's continuity theorem, the result follows. □

Remark. This theorem extends to $\mathbb{R}^d$ by using the next proposition, using the fact that $X_n \to X$ weakly in $\mathbb{R}^d$ if and only if $\langle X_n, v\rangle \to \langle X, v\rangle$ weakly in $\mathbb{R}$ for all $v \in \mathbb{R}^d$.

Definition. A random variable X in $\mathbb{R}^d$ is called a *Gaussian vector* if $\langle X_n, v\rangle$ are Gaussian for each $v \in \mathbb{R}^d$.

Proposition. Let X be a Gaussian vector in $\mathbb{R}^d$. Then $Z = AX + b$ is a Gaussian vector in $\mathbb{R}^m$ where A is an $m \times d$ matrix and $b \in \mathbb{R}^m$. Also, $X \in L^2(\mathbb{R}^d)$, and $\mu = \mathbb{E}[X]$ and $V = \operatorname{Cov}(X_i, X_j)$ exist and determine μ_X. The characteristic function is

$$\varphi_X(u) = e^{i\langle \mu, u\rangle - \frac{\langle u, Vu\rangle}{2}}$$

If V is invertible, then μ_X has a probability density function

$$f_X(x) = (2\pi)^{-\frac{d}{2}}(\det V)^{-\frac{1}{2}} \exp\left\{-\left\langle x-\mu, V^{-1}(x-\mu)\right\rangle\right\}$$

Subvectors $X_{(1)}, X_{(2)}$ of X are independent if and only if $\operatorname{Cov}(X_{(1)}, X_{(2)}) = 0$.

Proposition. Let $X_n \to X$ weakly in $\mathbb{R}^d$ as $n \to \infty$. Then,

(i) if $h : \mathbb{R}^d \to \mathbb{R}^k$ is continuous, then $h(X_n) \to h(X)$ weakly;

(ii) if $|X_n - Y_n| \to 0$ in probability, then $Y_n \to X$ weakly;

(iii) if $Y_n \to c$ in probability where c is constant on Ω, then $(X_n, Y_n) \to (X, c)$ weakly in $\mathbb{R}^d \times \mathbb{R}^d$.

Remark. Combining parts (iii) and (i), $X_n + Y_n \to X + c$ weakly if $Y_n \to c$ in probability. If $d = 1$, then in addition $X_n Y_n \to cX$ weakly.

Proof. Part (i). This follows from the fact that gh is continuous for any test function g.

Part (ii). Let $g : \mathbb{R}^d \to \mathbb{R}$ be bounded and Lipschitz continuous. Then

$$|\mathbb{E}\left[g(Y_n)\right] - \mathbb{E}\left[g(X)\right]| \leq \underbrace{|\mathbb{E}\left[g(X_n)\right] - \mathbb{E}\left[g(X)\right]|}_{<\frac{\varepsilon}{3}} + \mathbb{E}\left[|g(X_n) - g(Y_n)|\right]$$

where the bound on $\mathbb{E}\left[g(X_n)\right] - \mathbb{E}\left[g(X)\right]$ holds for sufficiently large n. Then the remaining term is upper bounded by

$$\begin{aligned} &\mathbb{E}\left[|g(X_n) - g(Y_n)|\right]\left(\mathbb{1}_{\left\{|X_n - Y_n| \leq \frac{\varepsilon}{3\|g\|_{\mathrm{Lip}}}\right\}} + \mathbb{1}_{\left\{|X_n - Y_n| > \frac{\varepsilon}{3\|g\|_{\mathrm{Lip}}}\right\}}\right) \\ &\leq \|g\|_{\mathrm{Lip}} \frac{\varepsilon}{3\|g\|_{\mathrm{Lip}}} + 2\|g\|_\infty \mathbb{P}\left(|X_n - Y_n| > \frac{\varepsilon}{3\|g\|_{\mathrm{Lip}}}\right) < \frac{2\varepsilon}{3} \end{aligned}$$

for sufficiently large n.

Part (iii). $|(X_n, c) - (X_n, Y_n)| = |Y_n - c| \to 0$ in probability. Also, $\mathbb{E}\left[g(X_n, c)\right] \to \mathbb{E}\left[g(X, c)\right]$ for all bounded continuous maps $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, so $(X_n, c) \to (X, c)$ weakly. Hence, by (ii), $(X_n, Y_n) \to (X, c)$ weakly. □

7. Ergodic theory

7.1. Laws of large numbers

Proposition. Let $(X_n)_{n\in\mathbb{N}}$ be independent and identically distributed random variables such that $\mathbb{E}[X_n] = 0$ and $\mathrm{Var}(X_n) = \sigma^2 < \infty$. Then $\frac{1}{n}\sum_{i=1}^n X_i \to 0$ in probability as $n \to \infty$.

Proof. By Chebyshev's inequality,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| > \varepsilon\right) \leq \frac{1}{n^2\varepsilon^2}\mathrm{Var}\left(\sum_{i=1}^n X_i\right) \leq \frac{\sigma^2}{n\varepsilon^2} \to 0$$

So $\frac{1}{n}\sum_{i=1}^n X_i \to \mathbb{E}[X_1]$ in probability. □

This is known as the weak law of large numbers. However, this result has several weaknesses, and we can provide stronger results.

Proposition. Let $(X_n)_{n\in\mathbb{N}}$ be independent random variables such that $\mathbb{E}[X_n] = \mu$ and $\mathbb{E}\left[X_n^4\right] \leq M$ for all n. Then $\frac{1}{n}\sum_{i=1}^n X_i \to \mu$ almost surely as $n \to \infty$.

Proof. Let $Y_n = X_n - \mu$. Then $\mathbb{E}[Y_n] = 0$, and $\mathbb{E}\left[Y_n^4\right] \leq 2^4\left(\mathbb{E}\left[X_n^4\right] + \mu^4\right) < \infty$. So we can assume $\mu = 0$. For distinct indices i, j, k, ℓ, by independence and the Cauchy–Schwarz inequality, we have

$$0 = \mathbb{E}\left[X_iX_jX_kX_\ell\right] = \mathbb{E}\left[X_i^2X_jX_j\right] = \mathbb{E}\left[X_i^3X_j\right]; \quad \mathbb{E}\left[X_i^2X_j^2\right] \leq \sqrt{\mathbb{E}\left[X_i^4\right]}\sqrt{\mathbb{E}\left[X_j^4\right]} \leq M$$

So we can compute

$$\mathbb{E}\left[\left(\sum_{i=1}^n X_i\right)^4\right] = \mathbb{E}\left[\sum_{i=1}^n X_i^4\right] + 6\mathbb{E}\left[\sum_{i<j} X_i^2X_j^2\right] \leq nM + 3n(n-1)M \leq 3n^2M$$

Let $S_n = \sum_{i=1}^n X_i$. Then,

$$\mathbb{E}\left[\sum_{n=1}^\infty \left(\frac{S_n}{n}\right)^4\right] \leq \sum_{n=1}^\infty \frac{1}{n^4}3n^2M < \infty$$

Hence $\sum_{n=1}^\infty \left(\frac{S_n}{n}\right)^4 < \infty$ almost surely. But then $\left(\frac{S_n}{n}\right)^4 \to 0$ almost surely, so $\frac{S_n}{n} \to 0$ almost surely. □

7.2. Invariants

Let $(E, \mathcal{E}, \mu)$ be a σ-finite measure space.

Definition. A measurable transformation $\Theta : E \to E$ is *measure-preserving* if $\mu(\Theta^{-1}(A)) = \mu(A)$ for all $A \in \mathcal{E}$.

In this case, for any integrable function $f \in L^1(\mu)$, we have $\int_E f \, \mathrm{d}\mu = \int_E f \circ \Theta \, \mathrm{d}\mu$.

Definition. A measurable map $f : E \to \mathbb{R}$ is called Θ-*invariant* if $f \circ \Theta = f$. A set $A \in \mathcal{E}$ is Θ-invariant if $\Theta^{-1}(A) = A$, or equivalently, $\mathbb{1}_A$ is Θ-invariant.

The collection $\mathcal{E}_\Theta$ of Θ-invariant sets forms a σ-algebra over E. A function $f : E \to \mathbb{R}$ is invariant if and only if f is $\mathcal{E}_\Theta$-measurable; this is a question on an example sheet.

Definition. Θ is called *ergodic* if the Θ-invariant sets A satisfy either $\mu(A) = 0$ or $\mu(E \setminus A) = 0$.

If f is Θ-invariant and Θ is ergodic, then one can show that f is constant almost everywhere on E.

Example. Consider $(E, \mathcal{E}) = ((0, 1], \mathcal{B})$ with the Lebesgue measure μ. The maps $\Theta_a(x) = x + a$ modulo 1 and $\Theta(x) = 2x$ modulo 1 are both measure-preserving, and ergodic unless $a \in \mathbb{Q}$. This is a question on an example sheet.

Lemma (maximal ergodic lemma)**.** Let $(E, \mathcal{E}, \mu)$ be a σ-finite measure space. Let $\Theta : E \to E$ be measure-preserving. For $f \in L^1(\mu)$, we define $S_0(f) = 0$ and $S_n(f) = \sum_{k=0}^{n-1} f \circ \Theta^k$. Let $S^\star = S^\star(f) = \sup_{n \geq 0} S_n(f)$. Then $\int_{\{S^\star > 0\}} f \, \mathrm{d}\mu \geq 0$.

Proof. Define $S_n^\star = \max_{k \leq n} S_k$. Then clearly $S_n^\star \uparrow S^\star$, and $S_k \leq S_n^\star$ for all $k \leq n$. Note that $S_{k+1} = S_k \circ \Theta + f \leq S_n^\star \circ \Theta + f$.

Define $A_n = \{S_n^\star > 0\}$, so $A_n \uparrow \{S^\star > 0\}$. On A_n, we have

$$S_n^\star = \max_{1 \leq k \leq n} S_k \leq \max_{0 \leq k \leq n} S_{k+1} \leq S_n^\star \circ \Theta + f$$

since $S_0 = 0$. We can integrate this inequality to find

$$\int_{A_n} S_n^\star \, \mathrm{d}\mu \leq \int_{A_n} S_n^\star \circ \Theta \, \mathrm{d}\mu + \int_{A_n} f \, \mathrm{d}\mu$$

On the complement A_n^c, we must have $S_n^\star = 0 \leq S_n^\star \circ \Theta$. Hence,

$$\int_E S_n^\star \, \mathrm{d}\mu \leq \int_E S_n^\star \circ \Theta \, \mathrm{d}\mu + \int_{A_n} f \, \mathrm{d}\mu$$

Since Θ is measure-preserving,

$$\int_E S_n^\star \, \mathrm{d}\mu \leq \int_E S_n^\star \, \mathrm{d}\mu + \int_{A_n} f \, \mathrm{d}\mu$$

so we obtain

$$\int_{A_n} f \, \mathrm{d}\mu \geq 0$$

Since $f \mathbb{1}_{A_n} \to f \mathbb{1}_{\{S^\star > 0\}}$ pointwise, and $|f \mathbb{1}_{A_n}| \leq |f| \in L^1(\mu)$, we can apply the dominated convergence theorem to show that

$$\int_{\{S^\star > 0\}} f \, \mathrm{d}\mu = \lim_{n \to \infty} \int_{A_n} f \, \mathrm{d}\mu \geq 0$$

as required. □

7.3. Ergodic theorems

Theorem (Birkhoff). Let $(E, \mathcal{E}, \mu)$ be a σ-finite measure space. Let $\Theta : E \to E$ be measure-preserving. For $f \in L^1(\mu)$, we define $S_0(f) = 0$ and $S_n(f) = \sum_{k=0}^{n-1} f \circ \Theta^k$. Then there exists a Θ-invariant integrable function $\overline{f} \in L^1(\mu)$ with $\mu\left(\left|\overline{f}\right|\right) \leq \mu(|f|)$ such that $\frac{S_n(f)}{n} \to \overline{f}$ almost everywhere.

The proof of Birkhoff's ergodic theorem is non-examinable.

Proof (non-examinable). Note that

$$\limsup_n \frac{S_n(f)}{n} = \limsup_n \frac{S_n(f) \circ \Theta}{n}$$

and the same holds for $\liminf_n$. Hence $\limsup_n \frac{S_n(f)}{n}$ and $\liminf_n \frac{S_n(f)}{n}$ are invariant functions. So they are $\mathcal{E}_\Theta$-measurable. Hence

$$D = D_{a,b} = \left\{ \liminf_n \frac{S_n(f)}{n} < a < b < \limsup_n \frac{S_n(f)}{n} \right\}$$

are measurable and invariant sets. Without loss of generality, let $b > 0$. Let $B \in \mathcal{E}$, where $B \subseteq D$ such that $\mu(B) < \infty$. Let $g = f - b \mathbb{1}_B \in L^1(\mu)$. Then,

$$S_n(g) = S_n(f) - b S_n(\mathbb{1}_B) \geq S_n(f) - bn$$

which is positive on D for some n by the definition of $\limsup_n$. We will apply the maximal ergodic lemma with $E = D$ and $\mu = \mu|_D$; Θ is still measure-preserving on this new measure since

$$\mu\Big|_D (A) = \mu(A \cap D) = \mu(\Theta^{-1}(A \cap D)) = \mu(\Theta^{-1}(A) \cap \Theta^{-1}(D)) = \mu(\Theta^{-1}(A) \cap D) = \mu\Big|_D (\Theta^{-1}(A))$$

Note that $\{S^\star > 0\} \subseteq D$ as we restrict our measure space to D, but by the previous inequality, $S^\star > 0$ on D. So $D = \{S^\star > 0\}$. Then the maximal ergodic lemma gives

$$0 \leq \int_{S^\star > 0} g \, \mathrm{d}\mu = \int_D g \, \mathrm{d}\mu = \int_D f \, \mathrm{d}\mu - b\mu(B)$$

Hence, $b\mu(B) \leq \int_D f \,\mathrm{d}\mu$. By σ-finiteness, this inequality extends to D; one can choose an approximating sequence $B_n \uparrow D$ where $\mu(B_n) < \infty$, then take limits to show $b\mu(D) = b\lim_n \mu(B_n) \leq \int_D f \,\mathrm{d}\mu$. Repeating the above argument for $-f$ and $-a$, we obtain $-a\mu(D) \leq \int_D -f \,\mathrm{d}\mu$. Combining these two inequalities gives

$$b\mu(D) \leq \int_D f \,\mathrm{d}\mu \leq a\mu(D)$$

But $a < b$, so $\mu(D) = 0$ or ∞, but f is integrable, so $\mu(D) = 0$. Now, define

$$\Delta = \left\{ \liminf_n \frac{S_n(f)}{n} < \limsup_n \frac{S_n(f)}{n} \right\} = \bigcup_{a<b\in\mathbb{Q}} D_{a,b}$$

By countable additivity,

$$\mu(\Delta) = \mu\left(\bigcup_{a<b\in\mathbb{Q}} D_{a,b} \right) = \sum_{a<b\in\mathbb{Q}} \mu(D_{a,b}) = 0$$

On Δ^c, $\frac{S_n}{n}$ converges in $[-\infty, \infty]$. We define the invariant function $\overline{f}$ by

$$\overline{f} = \begin{cases} \lim_n \frac{S_n}{n} & x \in \Delta^c \\ 0 & x \in \Delta \end{cases}$$

so $\frac{S_n}{f} \to \overline{f}$ almost everywhere as $n \to \infty$. Since $\mu(|f \circ \Theta^{n-1}|) = \mu(|f|)$, we have $\mu(|S_n|) \leq n\mu(|f|)$ and thus

$$\mu\left(\left|\overline{f}\right|\right) = \mu\left(\liminf_n \left|\frac{S_n}{n}\right|\right) \leq \liminf_n \mu\left(\left|\frac{S_n}{n}\right|\right) \leq \mu(|f|)$$

which is one of the results required by the theorem. In particular, $\mu\left(\left|\overline{f}\right|\right) < \infty$ so $\left|\overline{f}\right| < \infty$ almost everywhere. □

Theorem (von Neumann). Let $(E, \mathcal{E}, \mu)$ be a finite measure space (not σ-finite). Let $\Theta : E \to E$ be measure-preserving. Let $f \in L^p(E)$ with $1 \leq p < \infty$. Then $\frac{S_n(f)}{n} \to \overline{f}$ in L^p.

Proof. Since Θ is measure-preserving, we have

$$\left\| f \circ \Theta^i \right\|_p^p = \int_E |f|^p \circ \Theta^i \,\mathrm{d}\mu = \int_E |f|^p \,\mathrm{d}\mu = \int_E |f|^p \,\mathrm{d}\mu = \|f\|_p^p$$

Thus, by Minkowski's inequality, for all $f \in L^p$ we have

$$\left\| \frac{S_n(f)}{n} \right\|_p \leq \frac{1}{n} \sum_{i=0}^{n-1} \left\| f \circ \Theta^i \right\|_p = \|f\|_p$$

So $\frac{S_n(f)}{n}$ is a contraction in L^p. For each $K > 0$, we define $f_K = \max(\min(f, K), -K)$. Then

$$\|f - f_K\|_p^p = \int_E |f - f_K|^p \mathbb{1}_{|f|>K} \, \mathrm{d}\mu$$

Since $\mathbb{1}_{|f|>K}$ converges to zero pointwise, and $|f - f_K| \leq 2|f|^p \in L^1$, we find $\|f - f_K\|_p < \frac{\varepsilon}{3}$ by dominated convergence, for sufficiently large $K = K_\varepsilon$. As $|f_K| \leq K$, we have $\left|\frac{S_n(f_K)}{n}\right| \leq K$. Since μ is finite, $f_K \in L^1(\mu)$, so by Birkhoff's ergodic theorem, $\frac{S_n(f_K)}{n} \to \overline{f}_K$ almost everywhere for some invariant function $\overline{f}_K$. Note that $\overline{f}_k$ is bounded by K as $\frac{S_n(f_K)}{n}$ is bounded by K. By the bounded convergence theorem, we deduce that $\left\|\frac{S_n(f_K)}{n} - \overline{f}_K\right\| \to 0$ as $n \to \infty$. Further, this holds in L^p since

$$\left\|\frac{S_n(f_K)}{n} - \overline{f}_K\right\|_p \leq (2K)^{\frac{p-1}{p}} \left\|\frac{S_n(f_K)}{n} - \overline{f}_K\right\|_1 < \frac{\varepsilon}{3}$$

where the last inequality holds for sufficiently large n. Since μ is a finite measure, $L^p(\mu) \subseteq L^1(\mu)$, hence by Birkhoff's ergodic theorem, $\frac{S_n(f)}{n} \to \overline{f}$ almost everywhere as $f \to \infty$. Then, by the contraction property applied to $f - f_K$,

$$\begin{aligned}
\left\|\overline{f} - \overline{f}_K\right\|_p^p &= \int_E \left|\overline{f} - \overline{f}_K\right|^p \mathrm{d}\mu \\
&= \int_E \liminf_n \left|\frac{S_n(f) - S_n(f_K)}{n}\right|^p \mathrm{d}\mu \\
&\leq \liminf_n \int_E \left|\frac{S_n(f) - S_n(f_K)}{n}\right|^p \mathrm{d}\mu \\
&= \liminf_n \int_E \left|\frac{S_n(f - f_K)}{n}\right|^p \mathrm{d}\mu \\
&\leq \liminf_n \|f - f_K\|_p^p \\
&= \|f - f_K\|_p^p < \left(\frac{\varepsilon}{3}\right)^p
\end{aligned}$$

So in particular, $\overline{f} \in L^p$. Then by the triangle inequality,

$$\begin{aligned}
\left\|\frac{S_n(f)}{n} - \overline{f}\right\|_p &\leq \left\|\frac{S_n(f) - S_n(f_K)}{n}\right\|_p + \left\|\frac{S_n(f_K)}{n} - \overline{f}_K\right\|_p + \left\|\overline{f} - \overline{f}_K\right\|_p \\
&< \left\|\frac{S_n(f) - S_n(f_K)}{n}\right\|_p + \frac{2\varepsilon}{3} \\
&\leq \|f - f_K\|_p + \frac{2\varepsilon}{3} = \varepsilon
\end{aligned}$$

for sufficiently large n. □

7.4. Infinite product spaces

Let $E = \mathbb{R}^{\mathbb{N}} = \{x = (x_n)_{n\in\mathbb{N}}\}$ be the space of real sequences. Consider

$$\mathcal{C} = \left\{ A = \prod_{n=1}^{\infty} A_n \,\middle|\, A_n \in \mathcal{B}, \exists N \in \mathbb{N}, \forall n > N, A_n = \mathbb{R} \right\}$$

This forms a π-system, which generates the *cylindrical σ-algebra* $\sigma(\mathcal{C})$. One shows that $\sigma(\mathcal{C}) = \sigma(\{f_n \mid n \in \mathbb{N}\})$ where $f_n(x) = x_n$ are the coordinate projection functions on E. We can also show $\sigma(\mathcal{C}) = \mathcal{B}(\mathbb{R}^{\mathbb{N}})$ for the product topology. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with marginal distributions $\mu_{X_n} = m$ for all n; this exists by an earlier theorem. We define a map $X : \Omega \to E$ by $X(\omega)_n = X_n(\omega)$. This is $\mathcal{F}$–$\sigma(\mathcal{C})$ measurable, since for all $A \in \mathcal{C}$, we have

$$X^{-1}(A) = \{\omega \mid X_1(\omega) \in A_1, \dots, X_N(\omega) \in A_n\} = \bigcap_{n=1}^{N} X_n^{-1}(A_n) \in \mathcal{F}$$

We denote $\mu = \mathbb{P} \circ X^{-1}$, which is the unique product probability measure in $\mathbb{R}^{\mathbb{N}}$ satisfying

$$\begin{aligned}
\mu\left(\prod_{n=1}^{\infty} A_n\right) &= \lim_{N\to\infty} \mu\left(\prod_{n=1}^{N} A_n\right) \\
&= \lim_{N\to\infty} \mathbb{P}\left(X_1 \in A_1, \dots, X_N \in A_N\right) \\
&= \lim_{N\to\infty} \mathbb{P}\left(X_1 \in A_1\right) \cdots \mathbb{P}\left(X_N \in A_N\right) \\
&= \prod_{n=1}^{\infty} \mathbb{P}\left(X_n \in A_n\right) \\
&= \prod_{n=1}^{\infty} m(A_n)
\end{aligned}$$

Note that we need to use finiteness of N to exploit independence of the X_i. We call $(E, \mathcal{E}, \mu) = (\mathbb{R}^{\mathbb{N}}, \sigma(\mathcal{C}), m^{\mathbb{N}})$ the *canonical model* for an infinite sequence of random variables of law m.

Theorem. The shift map $\Theta : E \to E$ defined by $\Theta(x)_n = x_{n+1}$ is measure preserving and ergodic.

Proof. For $A \in \mathcal{C}$,

$$\begin{aligned}
\mu(A) &= \mathbb{P}\left(X_1 \in A_1, \dots, X_N \in A_N\right) \\
&= \mathbb{P}\left(X_1 \in A_1\right) \cdots \mathbb{P}\left(X_N \in A_N\right) \\
&= \prod_{n=1}^{N} m(A_n) \\
&= \mathbb{P}\left(X_2 \in A_1\right) \cdots \mathbb{P}\left(X_{N+1} \in A_N\right) \\
&= \mu(\Theta^{-1}(A))
\end{aligned}$$

so Θ is measure-preserving. Recall that the tail σ-algebra is defined by $\mathcal{T} = \bigcap_n \mathcal{T}_n$ where $\mathcal{T}_n = \sigma(\{X_k \mid k \geq n+1\})$. Note that for all $A \in \mathcal{C}$, we have

$$\Theta^{-n}(A) = \left\{x \in \mathbb{R}^{\mathbb{N}} \mid (x_{n+1}, x_{n+2}, \dots) \in A\right\} \in \mathcal{T}_n$$

Now, if A is invariant, $A = \Theta^{-n}(A) \in \mathcal{T}_n$ for all n, so $A \in \mathcal{T}$. By Kolmogorov's zero-one law, $\mu(A) = 0$ or $\mu(A) = 1$ as required for ergodicity. □

We can apply Birkhoff's ergodic theorem to Θ. If $f \in L^1(\mu)$, then $\frac{S_n(f)}{n} \to \overline{f} \in L^1(\mu)$ almost surely. Since $\overline{f}$ is invariant and Θ is ergodic, $\overline{f}$ is almost surely constant. By von Neumann's L^p-ergodic theorem, convergence holds in fact in L^1.

7.5. Strong law of large numbers

Theorem. Let $\int_{\mathbb{R}} |x| \,\mathrm{d}m(x) < \infty$, and let $\int_{\mathbb{R}} x \,\mathrm{d}m(x) = \nu$. Then

$$\mu\left(\left\{x \in \mathbb{R}^{\mathbb{N}} \,\middle|\, \frac{x_1 + x_2 + \dots + x_n}{n} \to \nu\right\}\right) = 1$$

Proof. Let $f(x) = x_1$. Then $f \in L^1(\mu)$, since $\int_E |f| \,\mathrm{d}\mu = \int_{\mathbb{R}} |x| \,\mathrm{d}m(x) < \infty$. So by Birkhoff's ergodic theorem,

$$\mu\left(\left\{\frac{x_1 + \dots + x_n}{n} \to \nu\right\}\right) = \mu\left(\left\{\frac{S_n(f)}{n} \to \overline{f}\right\}\right) = 1$$

where we also use von Neumann's ergodic theorem to deduce that

$$\overline{f} = \mu(\overline{f}) = \lim_n \mu\left(\frac{S_n(f)}{n}\right) = \frac{n}{n}\nu = \nu$$

□

Theorem (strong law of large numbers). Let $(X_n)_{n\in\mathbb{N}}$ be independent and identically distributed random variables such that $\mathbb{E}[|X_1|] < \infty$. Then $\frac{1}{n}\sum_{i=1}^n X_i \to \mathbb{E}[X]$ almost surely.

Proof. Inject X from Ω to $E = \mathbb{R}^{\mathbb{N}}$ as before, and notice that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \to \mathbb{E}[X]\right) = \mu\left(\left\{x \,\middle|\, \frac{x_1 + \dots + x_n}{n} \to \nu\right\}\right) = 1$$

□

Remark. The hypothesis $\mathbb{E}[|X|] < \infty$ cannot be weakened; we see on an example sheet that $\frac{1}{n}\sum_{i=1}^n X_i$ can exhibit various behaviours. Note that this notion of convergence is stronger than the weak convergence seen in the central limit theorem. The law of the iterated logarithm is that

$$\limsup_n \frac{X_1 + \dots + X_n}{\sqrt{2n \log\log n}} = 1$$

almost surely, and -1 for the limit inferior. In particular, the central limit theorem does not hold almost surely.

Corollary. By von Neumann's ergodic theorem, in the strong law of large numbers, we have $\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X]\right|\right] \to 0$ as $n \to \infty$.

III. Graph Theory

Lectured in Michaelmas 2022 by Dr. J. Sahasrabudhe

A graph is a set of vertices, each pair of which may be joined with an edge. The fact that graphs can be used to model any symmetric relation makes them widely applicable to various areas of mathematics such as knot theory.

We begin by studying various notions of connectivity of graphs, and then discuss planarity. The famous four-colour theorem states that a planar graph can be drawn using only four colours for the vertices, such that no two joined vertices share the same colour. While the proof of this theorem is extremely long, we prove the five-colour theorem and related results about graphs on other surfaces.

As graphs grow, we are interested in their asymptotic behaviour. For example, how many edges must there be in a graph before we can guarantee that there is a triangle? We study various properties of this form, and prove sufficient conditions to see certain behaviour in any given graph. We also use probability theory to provide bounds on how likely certain events in a random graph are to occur.

Contents

1. Introduction

1.1. Definitions

We use the notation $[n]$ for $\{1,\dots,n\}$. For a set X and $k \in \mathbb{N}$, we define $X^{(k)} = \{Y \subseteq X \mid |Y| = k\}$.

Definition. A *graph* is a pair (V,E), where V is a set of *vertices* and E is a set of *edges* where $E \subseteq V^{(2)}$. We use the notation $V(G)$ to denote the set of vertices and $E(G)$ to denote the set of edges, where $G = (V,E)$ is a graph. We define $|G| = |V(G)|$, and $e(G) = |E(G)|$.

Example. The complete graph on n vertices, denoted K_n, is the graph with $V = [n]$ and $E = V^{(2)}$.

Note that we sometimes use juxtaposition of names of vertices to denote an edge between them, so 13 represents the edge $\{1,3\}$.

Remark. Edges are undirected. There are no edges from a vertex to itself. Edges between vertices are unique if they exist. Most of the graphs covered in this course are finite.

Example. The empty graph on n vertices, denoted $\overline{K_n}$, is the graph with vertex set $V = [n]$ and $E = \varnothing$.

Example. The path of length n, denoted P_n, is the graph with vertex set $V = [n+1]$ and edge set $E = \{\{1,2\},\dots,\{n,n+1\}\}$.

Example. The cycle of length n, denoted C_n, is the graph with vertex set $V = [n]$ and edge set $E = \{\{1,2\},\dots,\{n-1,n\},\{n,1\}\}$.

Definition. Let G be a graph, $x \in V(G)$. The *neighbourhood* of x in G is

$$N_G(x) = \{y \in V(G) \mid \{x,y\} \in E(G)\}$$

If y is a neighbour of x, we write $x \sim y$.

Note that $\sim$ is irreflexive and not transitive in general.

Definition. The *degree* of a vertex $x \in V(G)$ is defined as $\deg x = |N(x)|$.

Definition. Let G, H be graphs. A *graph isomorphism* is a bijection $\varphi : V(G) \to V(H)$ such that $\{u,v\} \in E(G) \iff \{\varphi(u),\varphi(v)\} \in E(H)$.

Definition. We say H is a *subgraph* of G if $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$.

If G is a graph, and $xy \in E(G)$, we define $G - xy$ to be the graph $(V(G), E(G) \setminus \{xy\})$. Similarly, for $x, y \in V(G)$, we define $G + xy$ to be the graph $(V(G), E(G) \cup \{xy\})$.

Definition. Let $x, y \in V(G)$. A *walk* from x to y in G is a sequence of vertices $(x,\dots,y)$ such that each consecutive pair of elements of the sequence is connected by an edge in G. A *path* from x to y in G is a walk where all the vertices are disjoint.

Definition. A graph is *connected* if every pair of vertices is connected with a path.

The concatenation of two paths or walks P and P' is written PP'.

Remark. The concatenation of two walks is a walk. The concatenation of two paths is not necessarily a path, if the two paths share a vertex.

Proposition. If W is a x–y walk for $x \neq y$, W contains a x–y path, where 'contains' denotes a subsequence.

Proof. Let W' be the minimal x–y walk in W. This is a path, because if there were a repeated vertex, we could find a shorter path by eliminating the detour. □

Definition. We define the *distance* between two vertices, denoted $d(x, y)$, to be the shortest length of a path between x and y. If G is connected, this turns G into a metric space on its vertices.

1.2. Trees

Definition. A graph G is a *acyclic* if it does not contain a cycle C_k as a subgraph. A graph G is a *tree* if it is acyclic and connected.

Proposition. The following are equivalent.

(i) G is a tree (acyclic and connected).

(ii) G is *minimally connected*: G is connected and for all $xy \in E(G)$, $G - xy$ is not connected.

(iii) G is *maximally acyclic*: G is acyclic and for all $xy \notin E(G)$, $G + xy$ contains a cycle.

Proof. (i) implies (ii). Let $xy \in E(G)$. Suppose $G - xy$ were connected. Then there exists an x–y path P in $G - xy$. We can then close up the path P into a cycle in G by adding the edge xy. This contradicts the fact that G is acyclic.

(ii) implies (i). Suppose G has a cycle C. Let $xy \in E(C)$ be an edge in the cycle. We claim that $G - xy$ is connected. Let P be a u–v path in G. If P contains the edge xy, replace the use of this edge with the remainder of the cycle, traversed in the opposite direction. This yields a u–v walk in $G - xy$ which contains a u–v path.

(i) implies (iii). Let $xy \notin E(G)$. By connectedness, there exists an x–y path P in G. Hence, adding xy to $E(G)$, we obtain a cycle by concatenating P with xy.

(iii) implies (i). Suppose G is not connected. Then there exist $x \neq y$ such that there is no x–y path in G. Hence, adding xy to $E(G)$ cannot yield a cycle. □

Definition. Let T be a tree. A *leaf* of T is a vertex $v \in V(T)$ where $\deg(v) = 1$.

Definition. Let G be a graph, and $X \subseteq V(G)$. Then the *graph induced on* X, denoted $G[X]$ is the graph $(X, \{xy \in E(G) \mid x \in X, y \in X\})$. If $x \in G$, we define $G - x$ to be the graph $G[V(G) \setminus \{x\}]$.

Proposition. Let T be a tree where $|T| \geq 2$. Then T has a leaf.

Proof. Let $P = x_1, \dots, x_k$ be a longest possible path in T. $N(x_k) \subseteq \{x_1, \dots, x_{k-1}\}$ by maximality of P. If $x_i \sim x_k$ for any $1 \leq i \leq k-2$, we have a cycle, which is a contradiction. Hence $N(x_k) = \{x_{k-1}\}$, so x_k is a leaf. □

Remark. This proof actually demonstrates that any tree has at least two leaves, by considering x_1. We could alternatively have proven the lemma by taking a non-backtracking walk in G, which exists assuming no leaf exists; then, since $V(G)$ is finite, we must return to a point somewhere on the graph.

Proposition. Let T be a tree with $n \geq 1$ vertices. Then $|E(T)| = e(t) = n - 1$.

Proof. We prove this by induction on n. The $n = 1$ case is trivial. Now, assume that all trees with n vertices have $n-1$ edges, and suppose T has $n+1$ vertices. T has a leaf x. Then $T - x$ is a tree with n vertices since it is still connected, and hence has $n - 1$ edges. Since T has one more edge than $T - x$, namely the edge connecting the leaf x to $T - x$, T has n edges as required. □

Definition. Let G be a connected graph. Then a subgraph T of G is a *spanning tree* if $V(T) = V(G)$ and T is a tree.

Proposition. Every connected graph has a spanning tree.

Proof. Begin with G and remove edges of $E(G)$ such that the graph stays connected. When we can no longer remove edges, we must have a minimally connected subgraph of G, and hence a tree. □

1.3. Bipartite graphs

Definition. Let $G = (V, E)$ be a graph. G is *bipartite* if $V = A \cup B$ where $A \cap B = \varnothing$, such that all edges $(x, y) \in E$ satisfy $x \in A, y \in B$ or $x \in B, y \in A$.

The *complete bipartite graph* on n and m vertices, denoted $K_{n,m}$, is the bipartite graph with $|A| = n$, $|B| = m$ and with all possible edges.

Remark. Even cycles C_{2n} are bipartite, and odd cycles C_{2n+1} are not bipartite.

Definition. A *circuit* is a sequence $x_1, x_2, \dots, x_\ell, x_{\ell+1}$ where $x_i x_{i+1} \in E$ and $x_{\ell+1} = x_1$. In other words, a circuit is a closed walk. The *length* of this circuit is ℓ. A circuit is odd if its length is odd; a circuit is even if its length is even.

Proposition. Let C be an odd circuit in a graph G. Then C contains an odd cycle.

Proof. Let $x_1, \dots, x_\ell, x_1$ be an odd circuit. Either this is an odd cycle, or $x_i = x_j$ for $i < j$. Then $x_i, \dots, x_j$ is a circuit and $x_j, \dots, x_\ell, x_1, \dots, x_i$ is a circuit. Their lengths sum to ℓ, so one of them is odd. By induction, we can assume the odd circuit contains an odd cycle as required. □

Theorem. Let G be a graph. Then G is bipartite if and only if G does not contain an odd cycle.

Proof. If G contains an odd cycle, G is not bipartite because there exists a subgraph that is not bipartite. Suppose now that G contains no odd cycles. We may assume G is connected, because unions of disconnected bipartite graphs are bipartite. Let $x_0 \in V(G)$. Let $V_0 = \{x \in V(G) \mid d(x, x_0) \equiv 0 \bmod 2\}$ and $V_1 = \{x \in V(G) \mid d(x, x_0) \equiv 1 \bmod 2\}$. We show that this is a bipartition as required. Suppose $u, v \in V_i$ are connected. Then, u and v admit even (resp. odd) paths to x_0, so the circuit defined by the concatenation of these paths with the edge uv is an odd circuit, and hence contains an odd cycle. This contradicts our assumption. □

1.4. Planar graphs

Definition. A *plane graph* is a drawing of a graph in the plane, representing edges as piecewise linear functions, without edge crossings.

Definition. A graph G is *planar* if it can be drawn in the plane $\mathbb{R}^2$ with no edges crossing, so a graph is planar if it admits a plane graph representation.

Example. K_1, K_2, K_3, K_4 are planar. P_n is planar for $n \in \mathbb{N}$. $K_{n,2}$ is planar, by placing the vertices in the two-vertex set on either side of the other set.

Definition. Let G be a plane graph. One of the finitely many connected components of $\mathbb{R}^2 \setminus G$ is called a *face*. The boundary of a face F is the collection of vertices and edges in ∂f. Therefore, the boundary of any face in G is a subgraph of G.

Remark. The boundary of a face need not be (or contain) a cycle, and need not be connected. Two drawings of a graph can be fundamentally different.

Theorem (Euler)**.** Let G be a connected plane graph with F faces. Then $|V(G)|-|E(G)|+F = 2$.

Remark. The number of faces is uniquely determined by intrinsic properties of a graph, its number of vertices and edges.

Proof. We work by induction on the number of edges $E(G)$. In the case where $E(G) = 0$, we must have $V(G) = 1$ and $F = 1$ by connectedness. Suppose G is acyclic. Then by connectedness, G is a tree, so $V(G) = E(G) + 1$ and $F = 1$, satisfying Euler's formula. Now suppose G contains a cycle, and E be an edge in the cycle. Removing this edge, $G - E$ is connected, and has $|V(G)|$ vertices, $|E(G)| - 1$ edges, and $F - 1$ faces. By induction, Euler's formula holds in this case. □

Corollary. Let G be a planar graph where $|G| \geq 3$. Then $e(G) \leq 3|G| - 6$.

Proof. Consider a planar drawing of G. We may assume G is connected without loss of generality. Let F be a face, and let $\deg F$ be the number of edges that meet at F. Note that the degree of any face is at least 3, since $|G| \geq 3$. Since each edge occurs in at most two faces, $\sum_F \deg F \leq 2e(G)$. Hence, $3f \leq 2e(G)$, where f is the amount of faces. Using Euler's formula, $|G| - e(G) + f = 2 \implies 2(|G| - 2) \geq e(G)$. □

Remark. K_5 is not planar, because $e(K_5) = 10$ and $3|K_5| - 6 = 9$. $K_{3,3}$ does not violate this bound, but is not planar.

Corollary. Let G be a planar graph, $|G| \geq 4$ and there is no cycle of length 3. Then $e(G) \leq 2(|G| - 2)$.

Proof. The minimal degree of a face is 4, because a degree of 3 would imply there is a triangle since there are at least four vertices in the graph. Running the same argument, our bound becomes $e(G) \leq 2(|G| - 2)$, □

This shows that $K_{3,3}$ is not planar.

Definition. A *subdivision* of a graph G is a new graph where some of the edges are replaced by (disjoint) paths.

Remark. A subdivision of a non-planar graph is non-planar. In particular, if G contains a subdivision of $K_{3,3}$ or K_5, G is non-planar.

Theorem (Kuratowski). G is planar if and only if it contains no subdivision of $K_{3,3}$ or K_5.

2. Connectivity and matching

2.1. Matching in bipartite graphs

Definition. Let $G = (X \sqcup Y, E)$ be a bipartite graph. A *matching from X to Y* is a set of edges $E' \subseteq \{xy_x \mid x \in X, y_x \in Y\} = E$ such that the map $x \mapsto y_x$ is injective.

Definition. Let G be a graph, $A \subseteq V(G)$. We define $N_G(A) = \left\{\bigcup_{x\in A} N(x)\right\}$.

Theorem (Hall)**.** Let $G = (X \sqcup Y, E)$ be a bipartite graph. There exists a matching from X to Y if and only if *Hall's criterion* holds: that $|A| \leq |N(A)|$ for all $A \subseteq X$.

Proof. The forward direction is simple, by considering the image of the injective map $x \mapsto y_x : A \to N(A)$ for each subset $A \subseteq X$. Conversely, suppose Hall's criterion is satisfied. We apply induction on $|X|$. If $|X| = 1$, $N(X)$ is nonempty and so the proof is complete.

If there does not exist $\varnothing \neq A \subsetneq X$ such that $|N(A)| = |A|$, we have $|A| < |N(A)|$ for all $\varnothing \neq A \neq X$. Let $xy \in E$, and let $G' = G[X \setminus \{x\} \sqcup Y \setminus \{y\}]$. By induction, it suffices to show Hall's criterion holds for G'. If $B \subseteq X \setminus \{x\}$, we have

$$|N_{G'}(B)| \geq |N_G(B)| - 1 \geq |B|$$

as required.

However, suppose there exists such a set $\varnothing \neq A \subsetneq X$ with $|A| = |N(A)|$. Let $G_1 = G[A \sqcup N(A)]$ and $G_2 = G[X \setminus A \sqcup Y \setminus N(A)]$. G_1 satisfies Hall's criterion. Indeed, for $B \subseteq A$, $N_{G_1}(B) = N_G(B)$ as required. G_2 also satisfies Hall's criterion. Suppose $B \subseteq X \setminus A$, and consider $N_G(A \cup B)$. We have

$$|A| + |B| \leq |N_G(A \cup B)| = |N_G(A)| + \left|N_{G_2}(B)\right| \implies |B| \leq \left|N_{G_2}(B)\right|$$

Hence Hall's criterion is satisfied.

Then by induction on G_1 and G_2, the proof is complete. □

Definition. A *matching of deficiency d from X to Y* is a matching from $X' \subseteq X$ to Y where $|X'| + d = |X|$.

Theorem (defect Hall)**.** Let $G = (X \sqcup Y, E)$ be a bipartite graph. G contains a matching of deficiency $d \leq |X|$ if and only if $|A| \leq |N(A)| + d$ for all $A \subseteq X$.

Proof. The forward direction is again a simple proof. Let $G = (X \sqcup Y, E)$ be a graph such that $|A| \leq |N(A)| + d$ for all $A \subseteq X$. Let $G' = (X \sqcup (Y \cup \{z_1, \dots, z_d\}), E \cup E')$ where $E' = \{xz_i \mid x \in X, i \in \{1, \dots, d\}\}$. Hall's criterion on G' is satisfied, so there exists a matching. Deleting these new vertices $\{z_1, \dots, z_d\}$ and the edge set E', we construct a matching from X to Y of deficiency at most d. To construct a matching of deficiency precisely d, we can delete extra edges as required. □

Definition. The *maximum degree* $\Delta(G)$ (resp. *minimum degree* $\delta(G)$) of a graph G is the maximum (resp. minimum) degree of a vertex in G.

Definition. A graph is *regular* if all vertices have the same degree, or equivalently, $\delta(G) = \Delta(G)$. A graph is *k-regular* if $\delta(G) = \Delta(G) = k$.

Corollary. Let $G = (X \sqcup Y, E)$ be a k-regular bipartite graph and $k \geq 1$. Then there exists a matching from X to Y.

Proof. It suffices to show Hall's criterion holds. Let $A \subseteq X$. Then

$$e(G[A \cup N(A)]) = \sum_{x \in A} \deg x = k|A|; \quad e(G[A \cup N(A)]) = \sum_{x \in N(A)} \deg v \leq k|N(A)|$$

Hence $|A| \leq |N(A)|$. □

Example. Let Γ be a finite group, and let $H \leq \Gamma$. Let $L_1, \dots, L_n$ be the left cosets, and $R_1, \dots, R_n$ be the right cosets. We want to find $g_1, \dots, g_n$ such that $g_1 H, \dots, g_n H$ are the left cosets and $Hg_1, \dots, Hg_n$ are the right cosets.

Consider the graph $G = (\{L_1, \dots, L_n\} \sqcup \{R_1, \dots, R_n\}, E)$ where an edge lies between L_i and R_j if $L_i \cap R_j \neq \varnothing$. It suffices to find a matching in this graph, because then each edge in the matching implies the existence of a representative for both cosets. Let $A \subseteq \{L_1, \dots, L_n\}$, so $A = \left\{L_{i_1}, \dots, L_{i,k}\right\}$. Consider $\left|\bigcup_{j=1}^k L_{ij}\right| = k|H|$, but since $R_1, \dots, R_n$ partition Γ and have size $|H|$, at least k right cosets of H must intersect $\bigcup R_{ij}$. Hence Hall's criterion is satisfied.

2.2. Connectivity

Let $S \subseteq V(G)$. Then we define $G - S = G[V(G) \setminus S]$.

Definition. Let G be a graph, and $|G| \geq 1$. Then we define the *connectivity parameter* $\varkappa$ of G by

$$\varkappa = \min\{|S| \mid S \subseteq V(G), G - S \text{ is disconnected or a single vertex}\}$$

We say that G is k-connected if $k \leq \varkappa$. Hence G is k-connected if and only if for all sets S of at most $k-1$ vertices, $G - S$ is connected and not a single vertex.

Example. $\varkappa$(Petersen graph) $= 3$, because deleting any two vertices leaves the graph connected, but deleting the neighbourhood of any vertex disconnects the graph. $\varkappa(G) = 1$ if G is a tree. $\varkappa(C_n) = 2$ for $n \geq 3$. $\varkappa(K_n) = n - 1$.

Definition. Let G be a graph, and $a, b \in V(G)$. We say that the a–b paths $P_1, \dots, P_k$ are *disjoint* if $P_i \cap P_j = \{a, b\}$ for $i \neq j$.

Note that $\delta(G) \geq \varkappa(G)$. This follows because removing the neighbours of the vertex of minimum degree disconnects the graph or leaves it a single vertex. Also, we can easily see that $\varkappa(G-x) \geq \varkappa(G)-1$. Note that we can have $\varkappa(G-x) > \varkappa(G)$ by considering a 2-connected graph with an additional leaf.

Definition. Let G be a graph and $a \neq b \in V(G)$, where $a \nsim b$. We say that $S \subseteq V(G) \setminus \{a, b\}$ is a *a–b separator* if $G - S$ disconnects a and b.

Theorem (Menger, form 1)**.** Let G be a connected graph and $a \neq b \in V(G)$, where $a \nsim b$. The minimum size of an a–b separator is the maximum number of disjoint paths from a to b. Equivalently, if all a–b separators have size at least k, then there exist k disjoint a–b paths.

Proof. We write $\varkappa_{a,b}(G)$ for the minimum size of an a–b separator. Note that $\varkappa(G - x) \geq \varkappa(G) - 1$, and $\varkappa(G - xy) \geq \varkappa(G) - 1$. We also have the same properties for $\varkappa_{a,b}$.

Suppose the theorem does not hold, then there is a nonempty set of counterexamples. Let $\mathcal{G}$ be the set of counterexamples of smallest possible k, and let G be an element of $\mathcal{G}$ with the smallest possible amount of edges. Let S be a minimal a–b separator in G, so $|S| = k$. Note that the theorem is true for $k = 1$, so we may assume $k \geq 2$.

If $S \neq N(a)$ and $S \neq N(b)$, consider $G - S$. Then a, b lie in different connected components. Let A be the component containing a, and B be the component containing b. Define G_a to be the graph $G[A \cup S]$ together with a vertex c with edges to each $s \in S$. Similarly, define G_b to be the graph $G[B \cup S]$ together with a vertex c with edges to each $s \in S$.

Note that $\varkappa_{a,c}(G_a) \geq k$, because any a–c separator in G_a is an a–b separator in G, and $\varkappa_{b,c}(G_b) \geq k$ by symmetry. Note further that $e(G_a), e(G_b) < e(G)$; because $S \neq N(a)$ and $S \neq N(b)$, the amount of newly added edges is smaller than the amount of edges that must have been removed in each induced graph. Then by minimality of G, the G_a and G_b are not counterexamples to the theorem. Hence there exist disjoint a–c paths $P_1, \dots, P_k$ in G_a and disjoint c–b paths $Q_1, \dots, Q_k$ in G_b. Concatenating P_i with Q_i, we obtain k disjoint a–b paths in G. Then G is not a counterexample.

Now, suppose $S = N(a)$ without loss of generality. We claim that $N(a) \cap N(b) = \varnothing$. If there exists $x \in N(a) \cap N(b)$, then consider the graph $G - x$. We have $\varkappa_{a,b}(G - x) \geq k - 1$, so by minimality, there exist disjoint a–b paths $P_1, \dots, P_{k-1}$ in $G - x$. Adding the path a, x, b, which is disjoint from all others, we obtain k disjoint a–b paths, contradicting the assumption.

Let $a, x_1, \dots, x_\ell, b$ be a shortest a–b path. Note that $\ell \geq 2$ since $N(a) \cap N(b) = \varnothing$, and in particular, $x_2 \neq b$. Consider $G - x_1x_2$. We must have that $\varkappa_{a,b}(G - x_1x_2) \leq k - 1$, otherwise we have a smaller counterexample. Hence $\varkappa_{a,b}(G - x_1x_2) = k - 1$. Therefore there is an a–b separator $\tilde{S}$ with $|\tilde{S}| = k - 1$ in $G - x_1x_2$. We see that either $\tilde{S} \cup \{x_1\}$ or $\tilde{S} \cup \{x_2\}$ is a separator of size k in G, which is not equal to either $N(a)$ or $N(b)$. Then we can use the above construction to find the relevant contradiction. □

Corollary (Menger, form 2)**.** Let G be a connected graph with $|G| \geq 2$. Then G is k-connected if and only if all pairs of distinct vertices a, b admit k disjoint a–b paths.

Proof. Suppose all pairs of vertices a, b have k such paths. Suppose $G - S$ is disconnected, and a, b lie in different components of $G - S$. Note that $a \nsim b$, because there exists a separator for a and b. Then by assumption, there are k disjoint a–b paths, and so S must intersect each path. Therefore, $|S| \geq k$.

Now suppose G is k-connected. Let a, b be vertices in G. If $a \nsim b$, apply the first form of Menger's theorem. Conversely, consider $G - ab$. This graph is $k-1$-connected, so there are $k-1$ disjoint a–b by Menger's theorem. Adding the additional path a, b, we obtain k disjoint paths as required. □

2.3. Edge connectivity

Definition. Let G be a graph. Then $\lambda(G) = \min\{|W| \mid W \subseteq E(G), G - W \text{ disconnected}\}$ is the smallest amount of edges that can be deleted to disconnect G. We say that G is *k-edge connected* if $k \leq \lambda(G)$.

Example. Let C_n be the cycle on n vertices. The vertex connectivity $\varkappa$ and edge connectivity λ of this graph are both two.

Example. Consider a graph with two connected subgraphs K_n, but with one vertex in the intersection between the two. Then $\varkappa = 1$ by deleting the intersection vertex, but $\lambda(G) = n - 1$.

Definition. Paths $P_1, \dots, P_k$ are *edge-disjoint* if the edge sets are disjoint.

Theorem (Menger, edge version, form 1)**.** Let G be a connected graph, and $a \neq b$ be vertices. Then, if every $W \subseteq E(G)$ that separates a from b has size at least k, then there exist k edge-disjoint a–b paths.

Definition. Let G be a graph. The *line graph* of G, denoted $L(G)$, is the graph where $V(L(G)) = E(G)$ and $e, f \in E(G)$ are adjacent if they share an endpoint.

Proof. Let G' be the line graph of G, together with distinguished vertices a', b' that are connected to the edges incident to a and b respectively. Note that there is an a–b path in G if and only if there is an a'–b' path in G'. Thus, $W \subseteq V(G') \setminus \{a', b'\}$ is an a', b' separator if and only if $W \subseteq E(G)$ separates a from b. Therefore, $\varkappa_{a',b'}(G') \geq k$. By the first form of Menger's theorem on G', we can find k disjoint a'–b' paths $P_1, \dots, P_k$ in G'. These paths describe edge-disjoint a–b walks in G, which yield edge-disjoint a–b paths. □

Theorem (Menger, edge version, form 2)**.** Let G be a connected graph. Then $\lambda(G) \geq k$ if and only if all all pairs of vertices $a \neq b$ admit k edge-disjoint a–b paths.

Proof. If there exist k edge-disjoint paths between each pair of vertices, to separate any two vertices we must remove at least one edge from each of these k paths, so we must remove at least k edges. Conversely, if $\lambda(G) \geq k$, apply the above form of Menger's theorem. □

3. Colouring

3.1. Definition

Definition. A function $c \colon V(G) \to \{1, \dots, k\}$ is a *(proper) k-colouring* of a graph if $x \sim y \implies c(x) \neq c(y)$. The *chromatic number* of G, denoted $\chi(G)$, is the minimum k such that there exists a k-colouring of G.

Example. A path P_n has a 2-colouring. More generally, a graph is bipartite if and only if it has a 2-colouring. An even cycle has chromatic number 2, and an odd cycle has chromatic number 3. A tree has chromatic number 2. The complete graph on n vertices has chromatic number n.

Proposition. Let G be a graph. Then $\chi(G) \leq \Delta(G) + 1$.

Proof. Let $x_1, \dots, x_n$ be an ordering of the vertices of G. We create a colouring of the vertices by induction. Suppose $x_1, \dots, x_i$ have already been coloured, and we want to colour x_{i+1}. Since x_{i+1} has at most $\Delta(G)$ neighbours that have already been coloured, but we have $\Delta(G)+1$ available colours, there is a free colour that does not match any previous neighbours. Choose the smallest available colour. By induction we can colour the entire graph. □

Remark. This is sometimes known as a *greedy colouring*. The greedy colouring may produce a colouring which is suboptimal for a given graph; consider the path P_4 on the vertex set $\{1, 2, 3, 4\}$ but with the ordering $1, 4, 2, 3$: this gives a 3-colouring. The proposition above is sharp: the chromatic number of the complete graph is n, and its maximum degree is $n - 1$.

3.2. Colouring planar graphs

Proposition. Let G be planar. Then $\delta(G) \leq 5$.

Proof. The average degree of G, given by $n^{-1} \sum_{v \in V(G)} \deg v$, is exactly $2n^{-1}e(G)$. Since $e(G) \leq 3n - 6$, the average degree at most $6 - \frac{12}{n} < 6$, so $\delta(G) \leq 5$. □

Proposition (six-colour theorem)**.** Let G be planar. Then G admits a 6-colouring.

Proof. Apply induction on $|G|$. If $|G| \leq 6$, there admits a trivial 6-colouring. Let G be planar, and let $x \in V(G)$ have degree at most 5. By the inductive hypothesis, $G - x$ admits a 6-colouring. Since x has at most five neighbours, there is a free colour to use for x. □

Theorem (five-colour theorem)**.** Let G be planar. Then G admits a 5-colouring.

Proof. We apply induction on $|G|$. Clearly the theorem holds for $|G| \leq 5$. Suppose $|G| > 5$. Let $x \in V(G)$ be a vertex with degree at most five. Applying induction, there exists a 5-colouring of $G - x$. If the degree is four or lower, we can use the free colour to colour x, so suppose x has degree five. Let $N(x) = \{x_1, x_2, x_3, x_4, x_5\}$ arranged cyclically in the plane, and let the colour of x_i be i. Then without loss of generality, all the x_i must have different colours, since otherwise, we are done.

Suppose there exists no path from x_1 to x_3 in $G - x$ only along vertices coloured 1 and 3. In this case, let C be the component of G of vertices coloured 1 or 3 that contains x_1. This is the connected component of the subgraph of $G - x$ induced by the vertices coloured 1 and 3 that contains x_1. By assumption, x_3 is not in this component. Now, swap the colours 1 and 3 on C; this yields another 5-colouring of $G - x$. We can then extend this 5-colouring to x by colouring x with 1.

Now, suppose there exists no path from x_2 to x_4 in $G - x$ along vertices coloured 2 and 4. If so, we are done as above.

Suppose that there exists an x_1–x_3 path using only colours 1 and 3, and an x_2–x_4 path using only colours 2 and 4. Then since both paths lie in the plane and the vertices are arranged cyclically as above, they must cross. The intersection vertex is coloured either 1 or 3, and also either 2 or 4. This is a contradiction. □

Remark. Any planar graph admits a 4-colouring; this result is known as the *four-colour theorem*. The above method does not work, because when swapping the colours of a component, there is not a free colour to use for the newly added vertex. The four-colour theorem was eventually proven using a computer-aided search after reducing the problem to thousands of specific local configurations. The four-colour theorem is sharp; K_4 is planar.

3.3. Colouring non-planar graphs

Proposition. Let G be a connected graph, and $\delta(G) < \Delta(G)$. Then $\chi(G) \leq \Delta(G)$.

Proof. Order the vertices in G into $x_1, \dots, x_n$ such that $\deg x_n < \Delta(G)$, and x_{n-1} is adjacent to x_n, and also x_{n-2} is adjacent to one of x_n and x_{n-1} and so on. This is always possible since G is connected. This ordering has the property that all vertices have less than $\Delta(G)$ edges facing forward. So the greedy colouring gives a $\Delta(G)$-colouring. □

Theorem (Brooks)**.** Let G be a connected graph. If G is not an odd cycle or complete graph, $\chi(G) \leq \Delta(G)$.

Remark. We have shown above that $\chi(G) \leq \Delta(G) + 1$. This theorem then says that $\chi(G) = \Delta(G) + 1$ if and only if G is an odd cycle or a complete graph.

Proof. We apply induction on $|G|$. We can check that the theorem holds for $|G| \leq 3$. Note that we may assume that $\Delta(G) \geq 3$; otherwise, the graph is bipartite or an odd cycle.

We will show first that if G is 3-connected, the theorem holds. We give an ordering of $V(G)$. Let x_n be a vertex of degree $\Delta(G)$, and let $x_1, x_2 \in N(x)$ be non-adjacent vertices. This is possible; indeed, suppose we could not find such vertices. Then $\{x\} \cup N(x)$ is a complete graph, so $G = K_{\Delta(G)+1}$ by connectedness, contradicting our assumption. Now, consider

$G - \{x_1, x_2\}$. Since G is 3-connected, $G - \{x_1, x_2\}$ is connected. We can order the vertices in the same way as above, choosing $x_{n-1} \sim x_n$ and x_{n-2} a neighbour of x_{n-1} or x_n, and so on. Then the greedy algorithm produces the required colouring.

Now, we show that if $\varkappa(G) = 1$, the theorem holds. In this case, we have a separator of size one, so let $\{x\}$ be such a separator (we call x a *cut vertex*). Let $C_1, \dots, C_n$ be the connected components of $G - x$. By induction, we can colour $C_i \cup \{x\}$ for each i; they cannot be complete, by counting the number of edges of x in this graph. We can then permute the colours in each such colouring to make x the same colour. Then we can combine each colouring to produce a colouring of the entire graph.

Finally, we will consider the case when $\varkappa(G) = 2$. Let $S = \{x, y\}$ be a separator for G. Let $C_1, \dots, C_k$ be the components of $G - S$. Define the graphs $G_i = G[C_i \cup S] + xy$ for $i = 1, \dots, k$.

Suppose $\delta(G_i) < \Delta(G)$ for all i. In this case, the G_i can be coloured by induction as they are not complete graphs. Note that x, y get different colours since we have added the edge xy. Therefore, we can permute the colours, such that the colouring agrees on x, y for all G_i. These colourings can be combined into a $\Delta(G)$-colouring of G.

Now suppose without loss of generality that $\delta(G_1) = \Delta(G)$. In this case, $k = 2$, and

$$|N(x) \cap C_1| = \Delta(G) - 1 = |N(x) \cap C_1|; \quad |N(x) \cap C_2| = 1 = |N(y) \cap C_2|$$

Let x', y' be the neighbours of x, y in C_2. Now, note that $\tilde{S} = \{x, y'\}$ is a separator, and now $\delta(G_i) < \Delta(G)$ for all connected components, and we can use the proof from above. □

3.4. Chromatic polynomial

Definition. Let G be a graph. The *chromatic polynomial* of G is $P_G : \mathbb{Z}_{\geq 0} \to \mathbb{Z}_{\geq 0}$ where $P_G(t)$ is the number of t-colourings of G.

Remark. The minimum t for which $P_G(t) > 0$ the chromatic number.

Example. The chromatic polynomial on the empty graph on n vertices is given by $P_G(t) = t^n$.

The chromatic polynomial on the complete graph on n vertices is $P_G(t) = t(t-1)\dots(t-(n-1)) = n!\binom{t}{n}$.

For a path on n vertices, $P_G(t) = t(t-1)^{n-1}$. For any tree, colouring each leaf, removing it, then colouring the remainder inductively, $P_G(t) = t(t-1)^{|G|-1}$.

Definition. Let G be a graph, and $e = xy \in E(G)$. The *contraction* of G along e, denoted G/e, is the graph with vertices $V(G) \setminus \{x, y\} \cup \{a\}$ for a new variable a, and edges $E(G[V \setminus \{x, y\}]) \cup \{az \mid x \sim z\} \cup \{az \mid y \sim z\}$.

Proposition. Let G be a graph and $e \in E(G)$. Then $P_G(t) = P_{G-e}(t) - P_{G/e}(t)$.

Proof. Let $e = xy$. A t-colouring of $G - e$ where x, y are assigned different colours corresponds to a t-colouring of G, by simply adding the edge back. A t-colouring of $G - e$ where x, y are assigned the same colour corresponds to a t-colouring of G/e, by contracting the edge. □

Remark. The above proposition is known as a 'cut-fuse' relation.

Proposition. Let G be a graph. Then $P_G(t)$ is indeed a polynomial with degree $|G|$.

Proof. We apply induction on $e(G)$. If there are no edges in the graph, the graph is empty, and has chromatic polynomial $P_G(t) = t^{|G|}$. Otherwise, let $e \in E(G)$. By induction, $P_{G-e}(t)$ is a polynomial of degree $|G - e| = |G|$, and $P_{G/e}(t)$ is a polynomial of degree $|G/e| = |G| - 1$. Hence $P_G(t) = P_{G-e}(t) - P_{G/e}(t)$ is indeed a polynomial of the required degree. □

Proposition. Let G be a graph with n vertices and m edges. Then $P_G(t) = t^n - mt^{n-1} + p(t)$ where p is a polynomial of degree at most $n - 2$.

Proof. We apply induction on $e(G)$. If there are no edges, we have the empty graph, which has the required form. Otherwise, let $e \in E(G)$. Then

$$P_G(t) = P_{G-e}(t) - P_{G/e}(t) = \left(t^n - (m-1)t^{n-1} + \dots\right) + \left(t^{n-1} + \dots\right) = t^n - mt^{n-1} + \dots$$

as required. □

Remark. Other coefficients of the chromatic polynomial contain other information about the graph. For example, the t^{n-2} coefficient is exactly $\binom{e(G)}{2}$ – number of triangles in G.

If G is planar, $P_G\left(2 + \frac{1+\sqrt{5}}{2}\right) \neq 0$.

A result due to June Huh is that the coefficients $c_0, \dots, c_n$ of P_G are *log-concave*, so $c_i^2 > c_{i-1}c_{i+1}$.

3.5. Edge colouring

Definition. Let G be a graph. A *k-edge colouring* is a function $c : E(G) \to \{1, \dots, k\}$ such that if $c(e) \neq c(f)$ if e, f share an endpoint. The *edge chromatic number*, or the *chromatic index*, denoted $\chi'(G)$, is the minimum k such that there exists a k-edge colouring.

Remark. An edge colouring of G corresponds exactly to a vertex colouring of the line graph of G. In particular, $\chi'(G) = \chi(L(G))$. Note that not every graph can be realised as the line graph of some other graph.

Example. The edge chromatic number of an even cycle is 2. The edge chromatic number of an odd cycle is 3. This is because a cycle is its own line graph.

Example. We have $\Delta(G) \leq \chi'(G)$. If $x \in V(G)$ has degree $\Delta(G)$, all edges indicent to x must be given different colours. We may have $\Delta(G) < \chi'(G)$ for some graphs, such as C_3. The edge chromatic number of the Petersen graph is 4, but it is 3-regular.

We can show that $\chi'(G) \leq 2\Delta(G) - 1$ by the greedy colouring, considering how many vertices each edge can be connected to. χ' and χ can be very different, for instance, consider $\chi(K_{t,1}) = 2$ but $\chi'(K_{t,1}) = t$.

Given an edge colouring $c \colon E(G) \to \{1, \dots, k\}$, we define the *colour classes* as equivalence classes of colours: $C_i = \{e \in E(G) \mid c(e) = i\}$. Note that $(V(G), C_i \cup C_j)$ is the union of disjoint paths, even cycles, and isolated vertices. We say that the components of this graph are $\{i, j\}$-*components*.

Theorem (Vizing). Let G be a graph. Then $\chi'(G) = \Delta(G)$ or $\chi'(G) = \Delta(G) + 1$.

Proof. We prove this by induction on $|E(G)|$. It suffices to show there is a $\Delta(G) + 1$ colouring of any graph. If there are no edges, the graph can be 0-coloured, so $\chi'(G) = \Delta(G) = 0$ and so there is clearly a 1-colouring. For the inductive step, let G be a graph with $e(G) > 0$, and $xv \in E(G)$. Apply induction to $G - xv$ to obtain a $\Delta(G) + 1$ edge colouring.

Let $y \in V(G)$ and $c \in \{1, \dots, \Delta(G) + 1\}$. We say c is *missing* at y if no edge incident to y are coloured c. Note that there is a colour missing at every vertex since we have $\Delta(G) + 1$ different colours available.

Let c_0 be a colour missing at x. We define a sequence of vertices $v_1, \dots, v_k \in N(x)$ and corresponding colours $c_1, \dots, c_k$ such that c_i is missing at v_i. First, we set $v_1 = v$ and let c_1 be any colour missing at v. Then if v_i and c_i are defined, define v_{i+1} such that $c(xv_{i+1}) = c_i$, and define c_{i+1} to be any colour missing at v_{i+1}. This induction continues until either we find a colour missing at x or we repeat a colour.

Suppose $v_1, \dots, v_k$ are defined and c_k is missing at x. Then we can recolour xv_k with c_k. Now c_{k-1} is missing at x, so inductively, recolour xv_i with c_i. In particular, c_1 is missing at x, so we can colour xv_1 with c_1.

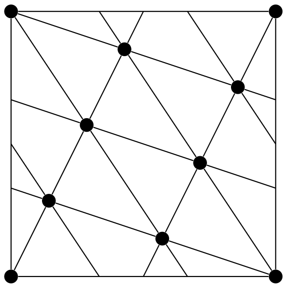
In the other case, suppose $c_k = c_i$ for $i < k$. Note that we may assume $i = 1$: uncolour xv_{i-1} and recolour xv_j with c_j for all $j < i$ as above. So $c_k = c_1$. If v_1 is not in the same $\{c_0, c_1\}$ component as x, we can swap the colours on the $\{c_0, c_1\}$ component containing v_1. Then c_0 is missing at v_1, and the colours of $xv_2, \dots, xv_k$ are unchanged. So we can colour xv_1 with c_0.

Now suppose x, v_1 are in the same $\{c_0, c_1\}$ component. If v_k is not in the same $\{c_0, c_1\}$ component as x, we can similarly swap the colours on the $\{c_0, c_1\}$ component containing v_k. So c_0 is missing at v_k and x, and so we can recolour xv_k to c_0, and inductively xv_i with c_i.

Now finally suppose x, v_1, v_k are all in the same $\{c_0, c_1\}$ component. So one of c_0, c_1 are missing at each of x, v_1, v_k. Since all $\{c_0, c_1\}$-components of the graph are disjoint paths, even cycles, or isolated vertices. So x, v_1, v_k are each endpoints of a path. But since paths only have two endpoints, this is a contradiction. □

3.6. Graphs on surfaces

We have seen that a planar graph has chromatic number $\chi(G) \leq 5$. Drawing graphs on other surfaces give different possible chromatic numbers. For instance, the complete graph on seven vertices K_7 can be drawn on a torus with no edge crossings.



Recall from IB Geometry that for any $g \in \mathbb{N}$, there is a *compact orientable surface of genus g* which is homeomorphic to a sphere with g 'handles' attached. The 2-sphere S^2 is a compact orientable surface of genus 0. The torus T^2 is a compact orientable surface of genus 1.

We have already seen that for a connected planar graph G with f faces, we have $|G| - e(G) + f = 2$. For a disconnected planar graph, we can add edges to make G into a connected graph. Hence, any planar graph with f faces satisfies $|G| - e(G) + f \leq 2$. In general, on the compact orientable surface of genus g, $|G| - e(G) + f \leq E$, where $E = 2 - 2g$ is the Euler characteristic of the surface. Due to results from IB Geometry, the equality holds for connected graphs, and then for any other graph, we can add edges to make it connected.

In particular, if $e(G) \geq 3$, then $3f \leq 2e(G)$ as usual. Therefore, $|G| - e(G) + \frac{2e(G)}{3} \geq E$, and so $e(G) \leq 3(|G| - E)$.

Theorem (Heawood). Let G be a graph drawn on a surface of Euler characteristic $E \leq 0$. Then

$$\chi(G) \leq H(E) = \left\lfloor \frac{7 + \sqrt{49 - 24E}}{2} \right\rfloor$$

Remark. Note that $H(2) = 4$, which would prove the four-colour theorem if not for the requirement that $E \leq 0$.

Proof. Let G be a graph drawn on a given surface with Euler characteristic E. Suppose its chromatic number is $\chi(G) = k$. Without loss of generality, we can choose a minimal such graph G with $\chi(G) = k$.

Each vertex has degree at least $k - 1$. Indeed, suppose there was a vertex of degree less than $k - 1$. Then we could remove this vertex and all associated edges, and we would obtain a strictly smaller graph with chromatic number exactly k, contradicting minimality. Further, we have $|G| \geq k$, otherwise we could colour the graph with only $|G|$ colours contradicting the definition of the chromatic number.

Since $e(G) \leq 3(|G| - E)$, the sum of the degrees of the vertices is $2e(G) \leq 6(|G| - E)$. Hence, $\delta(G) \leq \frac{1}{|G|} 6(|G| - E) = 6 - 6\frac{E}{|G|}$. In particular,

$$k - 1 \leq \delta(G) \leq 6 - 6\frac{E}{|G|} \leq 6 - 6\frac{E}{k}$$

Note that this step requires the fact that $E \leq 0$. This gives the quadratic equation $k^2 - 7k + 6E \leq 0$. Then,

$$\left(k - \frac{7}{2}\right)^2 - \frac{49}{4} + 6E \leq 0 \implies k \leq \frac{7 + \sqrt{49 - 24E}}{2}$$

□

Remark. The inequality is sharp, since the complete graph $K_{H(E)}$ can be drawn on a surface of characteristic E. An example of this is drawing K_7 on the torus, as demonstrated above. However, this is a very difficult result to prove.

4. Extremal graph theory

4.1. Hamiltonian graphs

Definition. A graph is said to be *Hamiltonian* if it contains a cycle that contains all vertices. Such a cycle is called a *Hamilton cycle*.

Theorem. Let G be a graph on $n \geq 3$ vertices. Then if $\delta(G) \geq \frac{n}{2}$, G is Hamiltonian.

Remark. This theorem is sharp. If n is even, two disjoint $K_{\frac{n}{2}}$ cliques suffices for a counterexample, since $\delta(G) = \frac{n}{2} - 1$. If n is odd, we can take two $K_{\frac{n+1}{2}}$ cliques which intersect in a single vertex, giving $\delta(G) = \frac{n-1}{2}$.

Proof. First, note that G is connected. Indeed, if $x \nsim y$, $|N(x)|, |N(y)| \geq \frac{n}{2}$, but there are only $n-2$ remaining vertices in the graph. So by the pigeonhole principle, there is a path of length 2 between x and y.

Consider a path $x_1, \dots, x_\ell$ of maximum length, and suppose for a contradiction that there is no cycle in G of length ℓ. Observe that $N(x_1) \subseteq \{x_2, \dots, x_{\ell-1}\}$ by maximality, and $N(x_\ell) \subseteq \{x_2, \dots, x_{\ell-1}\}$ by symmetry. Define $N^-(x_1) = \{x_i \mid x_{i+1} \in N(x_1)\}$. Note that $|N^-(x_1) \cup N(x_\ell)| \leq \ell - 1 \leq n - 1$, but $|N^-(x_1)|, |N(x_\ell)| \geq \frac{n}{2}$. So there exists $x_i \in N^-(x_1) \cap N(x_\ell)$. So we can find a cycle $x_i, x_\ell, x_{\ell-1}, \dots, x_{i+1}, x_1, x_2, \dots, x_i$ of length ℓ. □

Remark. Note that there is not an interesting theorem of the form '$e(G) \geq k$ implies G is Hamiltonian', because K_{n-1} adjoined to a single vertex by one edge is not Hamiltonian.

4.2. Paths of a given length

Lemma. Let G be a graph on n vertices, and $n \geq 3$. Let $k < n$. If G is connected and $\delta(G) \geq \frac{k}{2}$, then G contains a path of length k.

Remark. We need the assumption $k < n$, otherwise K_n is a counterexample. We need the assumption that G is connected, otherwise a collection of $\frac{n}{k}$ disjoint graphs give a counterexample if $n \mid k$. The requirement that $\delta(G) \geq \frac{k}{2}$ is sharp, by considering collections of $K_{\frac{k+1}{2}}$ that all intersect in a single vertex.

Proof. Let $x_1, \dots, x_\ell$ be a path of maximum length in G. There is no cycle of length ℓ, because if $\ell = n$ we are done as $k < n$, and if $\ell < n$ we can use a cycle of length ℓ to build a path of length $\ell + 1$ by the same argument from the previous theorem: $N^-(x_1)$ and $N(x_\ell)$ must intersect and so we can build a longer path. □

Theorem. Let G be a graph on n vertices. Then if $e(G) > \frac{n(k-1)}{2}$, G contains a path of length k.

Remark. If $k \mid n$, a collection of $\frac{n}{k}$ disjoint K_k graphs shows that the theorem is sharp.

Proof. Note that if $k = 1$, the theorem clearly holds. Suppose $k \geq 2$, and apply induction on n. The case $n = 2$ holds vacuously. Suppose now we have a graph G on $n \geq 3$ vertices. First note that $\frac{n(k-1)}{2} < e(G) \leq \frac{n(n-1)}{2}$, so $k < n$.

We may assume G is connected without loss of generality, because if it is disconnected, we can apply induction to one of its connected components. Let $C_1, \dots, C_r$ be the components, and $|C_i| = n_i$. Since $\sum_{i=1}^r e(G[C_i]) = e(G) > \frac{n(k-1)}{2}$, we have $\sum_{i=1}^r \left(e(G[C_i]) - \frac{n_i(k-1)}{2}\right) > 0$, so one of the summands is positive. So there exists a connected component C_i such that $e(G[C_i]) > \frac{n_i(k-1)}{2}$, so we can apply induction to this graph to obtain a path of length k as required.

If $\delta(G) \geq \frac{k}{2}$, the proof is complete by the previous lemma. Otherwise, there exists a vertex x of degree less than $\frac{k}{2}$, so $\deg(x) \leq \frac{k-1}{2}$. Note that $e(G - x) > \frac{n(k-1)}{2} - \frac{k-1}{2} = \frac{(n-1)(k-1)}{2}$, so we can apply induction to $G - x$ to obtain a path of length k, completing the proof. □

4.3. Forcing triangles

Proposition (Jensen). Let $a < b$ be real numbers, and $f : [a, b] \to \mathbb{R}$ be a convex function. Let $x_1, \dots, x_n \in [a, b]$. Then, $f\left(\frac{1}{n} \sum_{i=1}^n x_i\right) \leq \frac{1}{n} \sum_{i=1}^n f(x_i)$.

Theorem (Mantel). Let G be a graph on n vertices, and $\frac{n^2}{4} < e(G)$. Then G contains a triangle.

Remark. The bipartite graph $K_{\frac{n}{2},\frac{n}{2}}$ contains no triangles, and has $\frac{n^2}{4}$ edges, so the above theorem is sharp.

Proof. Suppose the graph contains no triangle. We may assume that $n \geq 3$, otherwise there is nothing to prove. Let $x, y \in V(G)$ such that $x \sim y$. In particular, $\deg x + \deg y \leq n-2+2 = n$.

Then, since $x \mapsto x^2$ is convex,

$$\begin{aligned} n \cdot e(G) &\geq \sum_{x\sim y} (\deg x + \deg y) \\ &= \frac{1}{2} \sum_x \sum_y (\deg x + \deg y) \mathbb{1}_{x\sim y} \\ &= \sum_x \sum_y \deg x \mathbb{1}_{x\sim y} \\ &= \sum_x \deg x \sum_y \mathbb{1}_{x\sim y} \\ &= \sum_x (\deg x)^2 \\ &= n\left(\frac{1}{n}\sum_x (\deg x)^2\right) \\ &\geq n\left(\frac{1}{n}\sum_x (\deg x)\right)^2 \\ &= n\left(\frac{2e(G)}{n}\right)^2 \end{aligned}$$

So $e(G) \leq \frac{n^2}{4}$ as required. □

4.4. Forcing cliques

Definition. We say that a graph G is *r-partite* if there is a partition of V into r subsets such that no part contains an edge. Equivalently, G is r-colourable, so $\chi(G) \leq r$.

Definition. Given natural numbers $n_1, \dots, n_r$, define $K_{n_1,\dots,n_r}$ to be the complete r-partite graph with partitions of size $n_1, \dots, n_r$.

Observe that if $r \mid n$, the graph $K_{\frac{n}{r},\dots,\frac{n}{r}}$ is an r-partite graph with $\binom{r}{2}\frac{n^2}{r^2} = \left(1 - \frac{1}{r}\right)\frac{n^2}{2}$ edges.

Theorem (Turán, form 1)**.** Let G be a graph on n vertices, and $\left(1 - \frac{1}{r}\right)\frac{n^2}{2} < e(G)$ for $r \geq 1$. Then G contains a subgraph of the form K_{r+1}, so it has an $(r+1)$-clique.

Proof. Suppose that G has an $(r+1)$-clique. For a given r, we prove the result by induction on n, assuming the theorem holds for all lower values of r, then we can complete the proof by induction.

If $n \leq r$, the result clearly holds. Let G be a graph that contains no $(r+1)$-clique. Suppose $r \geq 2$, otherwise the result is trivial. Then we can find an r-clique by induction on r. Let K be such a clique. Then each vertex in $V(G) \setminus K$ have at most $r-1$ neighbours in K, otherwise, this would be an $(r+1)$-clique. So

$$e(G) \leq \binom{r}{2} + (r-1)(n-r) + e(G \setminus K) \leq \binom{r}{2} + (r-1)(n-r) + \left(1 - \frac{1}{r}\right)\frac{(n-1)^2}{2} = \left(1 - \frac{1}{r}\right)\frac{n^2}{2}$$

□

Remark. This is a generalisation of Mantel's theorem. If $r \mid n$, the theorem is sharp by considering the complete r-partite graph.

Definition. The *Turán graph* $T_{r,n}$ is the complete r-partite graph $K_{n_1,\dots,n_r}$ where $\sum_{i=1}^r n_i = n$ and $n_1, \dots, n_r$ differ by at most one.

Proposition. Let G be a r-partite graph on n vertices. Then $e(G) \leq e(T_{r,n})$.

Remark. Turán graphs maximise the number of edges among all r-partite graphs on n vertices.

Proof. Let G be an r-partite graph on n vertices with the maximum number of edges. This graph is complete, since if there is a missing edge, there is a graph with more edges. Let $G = K_{n_1,\dots,n_r}$. Suppose that $n_i - n_j \geq 2$, so G is not a Turán graph. Then consider the graph obtained by moving an edge from the part with n_i vertices to the part with n_j vertices. Then we gain a total of $(n_i - 1)$ edges, and remove n_j edges. But this is at least 1, so we have obtained a graph with more edges. □

Theorem (Turán, form 2). Let G be a graph on n vertices and $r \geq 2$. Then if G does not contain an $(r+1)$-clique, $e(G) \leq e(T_{r,n})$.

Proof. We will transform a graph G into a complete r-partite graph without decreasing the number of edges. Then, since the Turán graph maximises the amount of edges for such a graph, the result follows.

Let $V(G) = \{1, \dots, n\}$. Let $\alpha_1, \dots, \alpha_r > 0$ be numbers that are linearly independent over $\mathbb{Q}$. For $S \subseteq V(G)$, define $\mu(S) = \sum_{i \in S} \alpha_i$.

If H is a graph on n vertices, we define the transformation of H, denoted $T(H)$, as follows. Let x, y be a pair of vertices maximising $\mu(\{x,y\})$ (to break any ties) such that $N(x) \neq N(y)$ and $x \nsim y$, and also either $\deg x > \deg y$ or both $\deg x = \deg y$ and $\mu(N(x)) > \mu(N(y))$. Now define $T(H)$ to be $H - y$ along with a new vertex x' with $N(x') = N(x)$.

We first show that if H does not contain a K_{r+1}, then $T(H)$ also does not contain a K_{r+1}. Suppose that our new graph H' contains a clique K isomorphic to K_{r+1}. We must have that x' lies inside this clique, because all other vertices remain the same. We know $x \notin K$ since $x \nsim x'$. Then $K \setminus \{x'\} \cup \{x\}$ must be an $(r+1)$-clique in H, which is a contradiction.

Now, consider the sequence $G, T(G), T(T(G)), \dots$, iteratively applying the transformation T. We will now show that this sequence $(T^{(n)}(G))_n$ eventually stabilises. This is because $e(T(H)) \geq e(H)$, so $(e(T^{(n)}(G)))_n$ is an increasing sequence of integers which is bounded above by $\binom{n}{2}$. Note that $\sum_{1\leq x\leq n}\mu(N_{T^{(i)}(G)}(x))$ is also an increasing sequence, but since there are only finitely many possible values for this sum, it must also stabilise. Therefore, at some point, the transformation T will do nothing more to our graph. Let G_∞ be the limiting graph in the sequence $(T^{(n)}(G))_n$.

We will show that G_∞ is a complete k-partite graph for some k. Let $k = \chi(G_\infty)$, and c be a k-colouring of G_∞. We write $V(G_\infty) = C_1 \cup \dots \cup C_k$ where C_i is the colour class of vertices with colour i. Note that if $x, y \in C_i$, we have $x \nsim y$, so $N(x) = N(y)$, otherwise the transformation T would have manipulated the neighbourhoods to be equal. Now let $x \in C_i, y \in C_j$ for $i \neq j$. Suppose $x \nsim y$. Then $x' \nsim y'$ for all other $x' \in C_i$ and $y' \in C_j$, so C_i and C_j have no edges between them. But then by merging C_i and C_j, we obtain a more optimal colouring, contradicting our assumption that $k = \chi(G_\infty)$. So G_∞ is a complete k-partite graph for some k.

Since G_∞ does not contain a K_{r+1}, we have $k \leq r$. By the previous proposition, $e(G_\infty) \leq e(T_{r,n})$, and $e(G) \leq e(G_\infty)$ since $e(H) \leq e(T(H))$ for all H. □

4.5. The Zarankiewicz problem

Definition. The *Zarankiewicz number* $Z(n, t)$ is the maximum number of edges in a bipartite graph $G = (X \sqcup Y, E)$ with $|X| = |Y| = n$ such that G does not contain $K_{t,t}$.

Lemma. Let $t \in \mathbb{N}$, and $t \geq 2$. Define the function $f_t(x) = \frac{x(x-1)\dots(x-t+1)}{t!}$. Then $f_t(x)$ is convex for $x \geq t - 1$.

Proof. Let $s = x - t + 1$, so $f_t(x) = \frac{(s+t-1)(s+t-2)\dots s}{t!}$. This is a polynomial with nonnegative coefficients. Hence it is convex for $s \geq 0$, since $f''(s) \geq 0$. □

Theorem. Let $t \geq 2$. Then $Z(n, t) \leq t^{\frac{1}{t}} n^{2-\frac{1}{t}} + tn$.

Remark. In particular, as n increases, $Z(n, t)$ is eventually lower bounded by $2n^{2-\frac{1}{t}}$.

Proof. Note that we may assume that $\deg y \geq t - 1$ for all $y \in Y$. If $\deg y < t - 1$, we can add an edge and preserve the property that G contains no $K_{t,t}$.

Let $x_1, \dots, x_t \in X$ be distinct vertices. Then $|N(x_1) \cap \dots \cap N(x_t)| \leq t-1$, otherwise we have a $K_{t,t}$. Now, applying Jensen's inequality,

$$\begin{aligned}
(t-1)\binom{n}{t} &\geq \sum_{x_1,\dots,x_t \text{ distinct}} |N(x_1) \cap \dots \cap N(x_t)| \\
&= \sum_{x_1,\dots,x_t \text{ distinct}} \sum_y \mathbb{1}_{y \sim x_1} \cdots \mathbb{1}_{y \sim x_t} \\
&= \sum_y \sum_{x_1,\dots,x_t \text{ distinct}} \mathbb{1}_{y \sim x_1} \cdots \mathbb{1}_{y \sim x_t} \\
&= \sum_y \binom{\deg y}{t} \\
&= n\left(\frac{1}{n}\sum_y \binom{\deg y}{t}\right) \\
&\geq n\binom{\overline{d}}{t}
\end{aligned}$$

where $\overline{d} = \frac{e(G)}{n}$ (since we are in a bipartite graph), using the fact that $\deg y \geq t-1$, so $x \mapsto \binom{x}{t}$ is convex. So

$$\begin{aligned}
(t-1)\binom{n}{t} &\geq n\binom{\overline{d}}{t} \\
\frac{tn^t}{t!} &\geq \frac{n(\overline{d}-t)^t}{t!} \\
tn^t &\geq n(\overline{d}-t)^t \\
t^{\frac{1}{t}}n^{1-\frac{1}{t}} &\geq \overline{d}-t \\
t^{\frac{1}{t}}n^{1-\frac{1}{t}} &\geq \frac{e(G)}{n}-t \\
e(G) &\leq t^{\frac{1}{t}}n^{2-\frac{1}{t}}+tn
\end{aligned}$$

□

Remark. If $t = 2$, then it is known that $Z(n,t) \geq cn^{\frac{3}{2}}$ for some contant $c > 0$. If $t = 3$, $Z(n,t) \geq cn^{\frac{5}{3}}$. This is an open problem for $t = 4$.

4.6. Erdős–Stone theorem

Definition. Let H be a fixed graph, and $n \in \mathbb{N}$. Then we define the *extremal number* $\operatorname{ex}(n,H) = \max\{e(G) \mid |G| = n, G \text{ contains no copy of } H\}$.

Example. $\operatorname{ex}(n,K_{r+1}) = e(T_{r,n}) \leq \left(1-\frac{1}{r}\right)\frac{n^2}{2}$. $\operatorname{ex}(n,P_k) = \frac{n(k-1)}{2}$. $\operatorname{ex}(n,K_{t,t}) \leq 2n^{2-\frac{1}{t}}+tn$.

Theorem. Let H be a fixed nonempty graph. Then

$$\lim_{n\to\infty} \frac{\mathrm{ex}(n,H)}{\binom{n}{2}} = 1 - \frac{1}{\chi(H)-1}$$

Remark. If $\chi(H) \geq 3$, this determines the leading order term in the function $\mathrm{ex}(n,H)$ for large n. If $\chi(H) = 2$, this theorem implies that $\frac{\mathrm{ex}(n,H)}{n^2} \to 0$. But in this case, $H \subseteq K_{t,t}$, and we already know (almost) that $\mathrm{ex}(n,H) \leq cn^{2-\frac{1}{t}}$, which implies the result from the Erdős–Stone theorem. It is easy to see that $\mathrm{ex}(n,H) \geq \left(1 - \frac{1}{\chi(H)-1}\right)\frac{n^2}{2}$, since H is not contained in any $T_{(\chi(H)-1),n}$.

5. Ramsey theory

5.1. Ramsey's theorem

Macroscopically, theorems in Ramsey theory are of the form 'complete disorder in sufficiently large systems is impossible'.

Proposition. Let c be a 2-edge (not proper) colouring of K_6. Then there exists a monochromatic triangle K_3; there exists a subgraph induced on three vertices where all edges have the same colour.

Proof. Suppose our colours are red and blue. Let $x \in V(K_6)$. Without loss of generality, x has three neighbours y_1, y_2, y_3 coloured red. Then the edges between the y_i cannot be coloured red. So they must all be coloured blue, but then this forms a blue triangle. □

Definition. Let $s \geq 2$. Then the *sth Ramsey number*, denoted $R(s)$, is the minimal n such that every 2-edge colouring of K_n contains a monochromatic K_s.

It is not clear *a priori* that such numbers indeed exist.

Definition. Let $s, t \geq 2$. We define $R(s,t)$ be the minimal n such that every 2-edge colouring of K_n contains either a red K_s or a blue K_t.

Remark. $R(s,t)$ is symmetric, and $R(s) = R(s,s)$. Note that $R(2,t)$ is the minimal n that contains a red edge or K_t, so $R(2,t) = t$. We showed above that $R(3,3) = R(3) \leq 6$, and in fact this is an equality by demonstrating a 2-edge colouring of K_5 containing no monochromatic triangle.

Theorem (Ramsey). For all s, t, the Ramsey number $R(s,t)$ exists, and $R(s,t) \leq R(s-1,t) + R(s,t-1)$.

Proof. Apply induction on $s+t$. For $s, t \leq 2$, the result holds. Now suppose $s, t > 2$, and let $a = R(s-1,t), b = R(s,t-1)$. Let $n = a + b = R(s-1,t) + R(s,t-1)$, and consider the complete graph K_n. Let $c : E(K_n) \to \{\text{red, blue}\}$ be a given colouring.

Let $x \in K_n$, and let $N_r(x)$ be the red neighbourhood and $N_b(x)$ be the blue neighbourhood. Suppose that $|N_r(x)| \geq a$. In this case, $N_r(x)$ contains either a red K_{s-1}, in which case $N_r(x) \cup \{x\}$ is a red K_s in K_n; or a blue K_t, in which case we are already done. Now suppose $|N_b(x)| \geq b$. Then $N_b(x)$ contains either a red K_s in which case we are done; or it contains a blue K_{t-1}, in which case $N_b(x) \cup \{x\}$ is a blue K_t in K_n as required. Suppose that neither of these cases occur, so $|N_r(x)| \leq a - 1$ and $|N_b(x)| \leq b - 1$, so $|N(x)| \leq a + b - 2$, which is a contradiction since the graph is complete. □

Corollary. For all s, the Ramsey number $R(s)$ exists.

Definition. $R_k(s_1, \dots, s_k)$ is the minimal n such that every k-edge colouring of K_n contains a K_{s_i} coloured i for some i.

Theorem (multicoloured Ramsey's theorem). For $s_1, \dots, s_k$ for $k \geq 2$, then $R_k(s_1, \dots, s_k)$ exists.

Proof. We will show by induction on k that $R_k(s_1, \dots, s_k) \leq R(s_1, R_{k-1}(s_2, \dots, s_k)) = n$. Let c be a k-colouring of K_n. Apply the two-colour version of Ramsey's theorem to obtain either a K_{s_1} coloured 1, or a $K_{R_{k-1}(s_2,\dots,s_k)}$ coloured in any combination of $2, \dots, k$. If we have a K_{s_1} coloured 1, we are done. Otherwise, apply induction to obtain an edge colouring of $K_{R_{k-1}(s_2,\dots,s_k)}$ to obtain a K_{s_i} coloured i for some $i \geq 2$. □

Remark. We have seen $R(3) = 6$. There are very few known Ramsey numbers. $R(4) = 18$, but $R(5)$ is unknown.

5.2. Infinite graphs and larger sets

Theorem. Let c be a 2-colouring of the countably infinite complete graph, so $c \colon \mathbb{N}^{(2)} \to \{\text{red}, \text{blue}\}$. Then there exists an infinite set $X \subseteq \mathbb{N}$ which is monochromatic, so $X^{(2)}$ is coloured either entirely red or entirely blue.

Remark. The finite version of Ramsey's theorem cannot be applied here; we can create arbitrarily large cliques, but we do not know if such cliques connect into an infinite set.

Proof. We construct a sequence $x_1, x_2, \dots$ inductively as follows. Let $x_1 \in \mathbb{N}$ be arbitrary. x_1 has either an infinite red neighbourhood or an infinite blue neighbourhood. We define S_1 to be the red neighbourhood of x_1 if it is infinite, or the blue neighbourhood otherwise, so S_1 is infinite. Now let $x_2 \in S_1$. Now, x_2 has either an infinite red neighbourhood in S_1 or an infinite blue neighbourhood in S_1, so we can define S_2 to be one of these that is infinite, and proceed inductively.

For each i, all edges $x_i \sim x_j$ where $i < j$ have the same colour by construction. Label a vertex red if all its forward-facing edges are red, and label an edge blue if all its forward-facing edges are blue. Then there are either infinitely many red vertices or infinitely many blue vertices. Without loss of generality, suppose the set of red vertices X is infinite. Then all edges in X are coloured red, so X is the infinite monochromatic set as required. □

Remark. We can easily construct a version of the above theorem for an arbitrary finite amount of colours, using the same idea as from the multiple-colour version of Ramsey's theorem in the finite case.

Example. It can be difficult to determine which colour has an infinite monochromatic clique. Suppose we colour ij with the maximal n such that $2^n \mid i + j$, modulo 2. The set $\left\{2^2, 2^4, 2^6, \dots\right\}$ is an example of an infinite monochromatic clique.

Suppose ij is coloured with the number of distinct prime factors of $i + j$, modulo 2. The colour of the infinite clique is not known.

Remark. It is possible to deduce the existence of $R(s, t)$ from the infinite version.

Theorem. Let c be a 2-colouring of the set of r-sets of $\mathbb{N}$, so $c : \mathbb{N}^{(r)} \to \{\text{red}, \text{blue}\}$. Then there exists an infinite set $X \subseteq \mathbb{N}$ such that $X^{(r)}$ is monochromatic.

Proof. Apply induction on r. If $r = 2$, we fall back to the previous theorem. We define a sequence $x_1, x_2, \dots$ and a sequence of infinite sets $S_1, S_2, \dots$ by the following procedure. We start by choosing x_1 arbitrarily. Now, consider the colouring $c_{x_1}(F) = c(\{x_1\} \cup F)$ for $F \in (\mathbb{N} \setminus \{x_1\})^{(r-1)}$. By induction, there exists a set $S_1 \subseteq \mathbb{N} \setminus \{x_1\}$ that is infinite and $S_1^{(r-1)}$ is monochromatic with respect to the colouring c_{x_1}. Now we choose $x_2 \in S_1$, and proceed inductively.

The sequence $x_1, x_2, \dots$ has the property that $F_i = \left\{\{x_{i_1}, \dots, x_{i_r}\} \mid i_1 < \dots < i_r\right\}$ are monochromatic for each i. But there are either infinitely many red-coloured x_i or infinitely many blue-coloured x_i. Let X be one of these infinite sets, then $X^{(r)}$ is monochromatic. □

We can produce a similar version of this theorem for the finite case, along with an explicit inductively-defined bound.

Definition. Let $r \in \mathbb{N}$, and $s, t \geq 1$. We define the *r-set Ramsey number* $R^{(r)}(s,t)$ to be the minimal n such that for every 2-colouring of $\{1, \dots, n\}^{(r)}$, it contains either a set S with $|S| = s$ and $S^{(r)}$ are coloured red, or a set T with $|T| = t$ and $T^{(r)}$ are coloured blue.

Remark. $R^{(1)}(s,t) = s + t - 1$. $R^{(2)}(s,t) = R(s,t)$. $R^{(r)}(r,t) = t = R^{(r)}(t,r)$.

Theorem. For all $r, s, t \geq 1$, the number $R^{(r)}(s,t)$ exists.

Proof. Apply induction on r, and then induction on $s+t$. If $s \leq r$ or $t \leq r$, we are done, since $R^{(r)}(r,t) = t$. We claim that $R^{(r)}(s,t) \leq R^{(r-1)}(R^{(r)}(s-1,t) + R^{(r)}(s,t-1)) + 1 = N$.

Consider a 2-coloured set $\{1, \dots, n\}^{(r)}$ where $n \geq N$. Choose a vertex $x \in \{1, \dots, n\}$. Consider the colouring $c_x(F) = c(\{x\} \cup F)$ where $F \in (\{1, \dots, n\} \setminus \{x\})^{(r-1)}$. Applying induction on r, we have a set S_1 such that $|S_1| = R^{(r)}(s-1,t)$ and $S_1^{(r-1)}$ is red, or there is a set S_2 with $|S_2| = R^{(r)}(s,t-1)$ and $S_2^{(r-1)}$ is blue. We consider the first case; the other is similar.

Apply the r-set version of Ramsey's theorem by induction to S_1 to find either a set $A \subseteq S_1$ with $|A| = s - 1$ and $A^{(r)}$ is coloured red (with respect to c), or a set $B \subseteq S_2$ with $|B| = t$ and $B^{(r)}$ is coloured blue. If B exists, we are done. If A exists, $A \cup \{x\}$ is coloured red and has size s as required. □

5.3. Upper bounds

Proposition. Let $s, t \geq 2$, we have $R(s,t) \leq \binom{s+t-2}{t-1}$. In particular, $R(s) = R(s,s) \leq 4^s$.

Proof. Apply induction on $s + t$. We know $R(s,2) = s = \binom{s+2-2}{2-1}$ as required. Suppose this holds for $R(s-1,t)$ and $R(s,t-1)$. We have already shown that $R(s,t) \leq R(s-1,t) + R(s,t-1)$.

So

$$R(s,t) \leq R(s-1,t) + R(s,t-1) \leq \binom{s+t-2}{s-2} + \binom{s+t-3}{s-1} = \binom{s+t-2}{s-1}$$

□

We are interested in bounding $R^{(r)}(s,t)$. Note that we have the bound $R^{(r)}(s,t) \leq R^{(r-1)}(R^{(r)}(s,t-1), R^{(r)}(s-1,t)) + 1$. Define $f_1(x) = 2x$, and recursively, $f_n(x) = f_{n-1}^x(x)$. Then $f_2(x) \sim 2^x$, and as n increases, f_n increases very rapidly. So our bound on $R^{(r)}(s,t)$ grows asymptotically on the order of $f_r(s+t)$.

5.4. Lower bounds

We can explicitly construct some lower bounds for $R(s)$.

Proposition. $R(s) > (s-1)^2$.

Proof. Consider the graph defined by $(s-1)$ disjoint K_{s-1} cliques, all of which are coloured blue, but all lines between cliques are coloured red. This graph has no monochromatic K_s. □

Theorem (Erdős). Let $s \geq 3$. Then $R(s) \geq 2^{\frac{s}{2}}$.

Proof. Consider $G = K_n$ for $n \leq 2^{\frac{s}{2}}$. For each edge e in G, we construct an independent Bernoulli random variable X_e with parameter $\frac{1}{2}$. If $X_e = 0$, we colour e red, and if $X_e = 1$, we colour e blue. Then

$$\begin{aligned}
\mathbb{P}\left(\text{colouring has a monochromatic } K_s\right) &= \mathbb{P}\left(\bigcup_{K \in \{1,\dots,n\}^{(s)}} \{K \text{ monochromatic}\}\right) \\
&\leq \sum_{K \in \{1,\dots,n\}^{(s)}} \mathbb{P}\left(K \text{ monochromatic}\right) \\
&= \sum_{K \in \{1,\dots,n\}^{(s)}} 2 \cdot 2^{-\binom{s}{2}} \\
&= \binom{n}{s} 2 \cdot 2^{-\binom{s}{2}} \\
&< \frac{n^s}{s!} 2 \cdot 2^{-\frac{s(s-1)}{2}} \\
&= 2\left(\frac{n}{(s!)^{\frac{1}{s}}} 2^{-\frac{s-1}{2}}\right)^s \\
&\leq 2\left(\frac{2^{\frac{1}{2}}}{(s!)^{\frac{1}{s}}}\right)^s
\end{aligned}$$

Note that $s! \geq 2^{\frac{s}{2}+1}$, so $(s!)^{\frac{1}{s}} \geq 2^{\frac{1}{2}+\frac{1}{s}}$.

$$\mathbb{P}\left(\text{colouring has a monochromatic } K_s\right) < 2\left(\frac{1}{2^{\frac{1}{s}}}\right)^s \leq 1$$

Since the probability is less than 1, there must exist a colouring that has no monochromatic K_s. □

Remark. We can think about this proof as follows. Consider the collection of $2^{\binom{n}{2}}$ colourings of K_n. Then for each clique, there are at most $2^{\binom{n}{2}} \cdot 2 \cdot 2^{-\binom{s}{2}}$ colourings where that clique is monochromatic. So the collection of all colourings where none of these cliques are monochromatic has at least as many elements as $2^{\binom{n}{2}} - \binom{n}{2} 2^{\binom{n}{2}} \cdot 2 \cdot 2^{-\binom{s}{2}}$. In general, however, a probabilistic interpretation is more powerful.

Remark. This proof is nonconstructive. It is a major open problem to explicitly construct colourings to show that $R(s) > (1+\varepsilon)^s$.

6. Random graphs

6.1. Lower bounds for Zarankiewicz numbers

Recall the Zarankiewicz numbers $Z(n,t)$, the maximum number of edges between a bipartite graph on (n,n) vertices, before a $K_{t,t}$ is forced. We have shown that $Z(n,t) \leq 2n^{2-\frac{1}{t}}$, but we have found no lower bound.

Theorem. Let $t \geq 2$. Then $Z(n,t) \geq \frac{1}{2}n^{2-\frac{2}{t+1}}$.

Proof excluding the $t+1$ *term.* Suppose we include each edge in the graph with probability p. Let Z be a random variable that counts the number of $K_{t,t}$ in the bipartite graph G on (n,n) vertices. Then

$$Z = \sum_{A \in X^{(t)}, B \in Y^{(t)}} \mathbb{1}(\text{all edges between } A \text{ and } B \text{ lie in } G)$$

We find

$$\mathbb{E}\left[Z\right] = \sum_{A \in X^{(t)}, B \in Y^{(t)}} \mathbb{P}\left(\text{all edges between } A \text{ and } B \text{ lie in } G\right) = \binom{n}{t}^2 p^{t^2} \leq \frac{n^{2t}}{4} p^{t^2} = \frac{1}{4}(n^2p^t)^t$$

So if $p = n^{-\frac{2}{t}}$, then our upper bound is at most $\frac{1}{4}$. Then $\mathbb{P}\left(X \geq 1\right) \leq \frac{1}{4}$ by Markov's inequality. Note that $\mathbb{E}\left[e(G)\right] = pa^2 = n^{2-\frac{2}{t}}$. So $\mathbb{P}\left(e(G) \leq \frac{pn^2}{2}\right) \leq \frac{1}{2}$. So with probability greater than $\frac{1}{4}$, we have $e(G) > \frac{1}{2}pn^2 = \frac{1}{2}n^{2-\frac{2}{t}}$ and G does not contain a $K_{t,t}$. □

Proof. Let $G = (X \sqcup Y, E)$ be a random bipartite graph with $|X| = |Y| = n$, such that $xy \in E$ with probability $p = n^{-\frac{2}{t+1}}$. Let $\widetilde{G}$ be the graph G with an edge removed from each $K_{t,t}$. By definition, $\widetilde{G}$ has no $K_{t,t}$. Note that $e(\widetilde{G}) \geq e(G) - (\text{amount of } K_{t,t} \text{ in } G)$. Taking expectations, $\mathbb{E}\left[e(\widetilde{G})\right] \geq \mathbb{E}\left[e(G)\right] - \mathbb{E}\left[\text{amount of } K_{t,t}\right]$. We have $\mathbb{E}\left[e(G)\right] = pn^2$, and the expected amount of $K_{t,t}$ subgraphs of G is $\binom{n}{t}^2 p^{t^2}$. Substituting in for p and approximating,

$$\mathbb{E}\left[e(\widetilde{G})\right] \geq n^{2-\frac{2}{t+1}} - \frac{n^{2t}}{2}p^{t^2}$$

Note that

$$n^{2t}p^{t^2} = (n^2p^t)^t = (n^2n^{\frac{-2t}{t+1}})^t = (n^{\frac{2(t+1)-2t}{t+1}})^t = n^{\frac{2t}{t+1}} = n^{2-\frac{2}{t+1}}$$

Hence

$$\mathbb{E}\left[e(\widetilde{G})\right] \geq \frac{1}{2}n^{2-\frac{2}{t+1}}$$

So there must exist a graph $\widetilde{G}$ with no $K_{t,t}$ and that has at least $\frac{1}{2}n^{2-\frac{2}{t+1}}$ edges. □

6.2. Girth

Definition. The *girth* of a graph is the length of the shortest cycle.

Proposition (Markov)**.** Let X be a nonnegative random variable. Then for all $t > 0$,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$

Proposition. Let G be a graph. Then $\chi(G) \geq \frac{|G|}{\alpha(G)}$, where $\alpha(G)$ is the size of the largest independent set (non-adjacent vertices) in G.

Proof. Let c be a colouring of G with $k = \chi(G)$ colours. Let C_i be the set of vertices coloured i. Then the C_i are each independent sets. We have $|G| = |C_1| + \dots + |C_k| \leq k\alpha(G) = \chi(G)\alpha(G)$. □

Theorem (Erdős)**.** For all $k, g \geq 3$, there exists a graph G with $\chi(G) \geq k$ and girth at least g.

Proof. Let G be a random graph on $\{1, \dots, n\}$ where each edge ij is included with probability $p = n^{-1+\frac{1}{g}}$. Let X_i be the random variable that counts the number of cycles in G of length i. Let $X = X_3 + \dots + X_{g-1}$. Now, note that $\mathbb{P}\left(X \geq \frac{n}{2}\right) \leq \frac{2}{n}\mathbb{E}[X]$.

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=3}^{g-1} \mathbb{E}[X_i] \\
&\leq \sum_{i=3}^{g-1} \frac{n(n-1)\dots(n-i+1)}{i} p^i \\
&\leq \sum_{i=3}^{g-1} (np)^i \\
&= \sum_{i=3}^{g-1} n^{\frac{i}{g}} \\
&\leq cn^{-\frac{1}{g}} < \frac{1}{2}
\end{aligned}$$

for a constant c. Now, let Y be the random variable counting the number of independent sets of $s = \frac{n}{2k}$ vertices (up to rounding).

$$\begin{aligned}\mathbb{P}(Y \geq 1) &\leq \mathbb{E}[Y] \\ &= \binom{n}{s}(1-p)^{\binom{s}{2}} \\ &\leq n^s e^{-p\binom{s}{2}} \\ &= \left(n^2 e^{-p(s-1)}\right)^{\frac{s}{2}} \\ &\leq \left(2n^2 e^{-\frac{n^{\frac{1}{g}}}{2k}}\right)^{\frac{s}{2}} \\ &< \frac{1}{2}\end{aligned}$$

for n sufficiently large. We have shown that G has at most $\frac{n}{2}$ cycles of length at most $g-1$ with probability at least $\frac{1}{2}$, and G has $\alpha(G) \leq \frac{n}{2k}$ with probability at least $\frac{1}{2}$. Hence there is a graph G with both properties. Let $\widetilde{G}$ be G with a vertex deleted from each cycle of length less than g. Then $\widetilde{G}$ has girth at least g. Further,

$$\chi(\widetilde{G}) \geq \frac{|\widetilde{G}|}{\alpha(\widetilde{G})} \geq \frac{\frac{n}{2}}{\alpha(G)} \geq \frac{\frac{n}{2}}{\frac{n}{2k}} = k$$

as required. □

6.3. Binomial random graphs

Definition. The *binomial random graph* on n vertices with parameter $p \in [0,1]$ is the probability space $G(n,p)$ on the graphs on n vertices, where each potential edge is included in the graph independently with probability p.

Let $(a_n), (b_n)$ be sequences of nonnegative numbers, and $b_n \neq 0$ for sufficiently large n. Then we write $a_n \ll b_n$ if $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$. Let X be the random variable that counts the number of triangles K_3 in some random graph $G \sim G(n,p)$. Then $\mathbb{E}[X] = \binom{n}{3}p^3$.

Note that if $p \ll \frac{1}{n}$, so $pn \to 0$, we have $\mathbb{E}[X] \leq n^3p^3 \to 0$. By Markov's inequality, $\mathbb{P}(K_3 \subset G) = \mathbb{P}(X \geq 1) \leq \mathbb{E}[X] \to 0$.

If $p \gg \frac{1}{n}$, so $pn \to \infty$, then we have $\mathbb{E}[X] \geq \frac{(n-3)^3}{6}p^3 \to \infty$. So asymptotically we have infinitely many triangles. We can also show that $\mathbb{P}(X \geq 1) \to 1$, but this does not follow immediately from the previous result.

Proposition (Chebyshev). Let X be a random variable, and let $t > 0$. Then

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}$$

Proposition (second moment method)**.** Let X be a random variable taking values in $\mathbb{N}$. Then

$$\mathbb{P}(X = 0) \leq \frac{\mathrm{Var}(X)}{(\mathbb{E}[X])^2}$$

Proof.

$$\mathbb{P}(X = 0) \leq \mathbb{P}(|X - \mathbb{E}[X]| \geq \mathbb{E}[X]) \leq \frac{\mathrm{Var}(X)}{(\mathbb{E}[X])^2}$$

□

Theorem. Let $G \sim G(n, p)$ be a binomial random graph. Then

$$\lim_{n\to\infty} \mathbb{P}(K_3 \subset G) = \begin{cases} 0 & p \ll \frac{1}{n} \\ 1 & p \gg \frac{1}{n} \end{cases}$$

Proof. Let X be the random variable counting the triangles in G. If $p \ll \frac{1}{n}$, then $\mathbb{E}[X] \to 0$ so $\mathbb{P}(X \geq 1) \to 0$. Now suppose $p \gg \frac{1}{n}$. Let $p \gg \frac{1}{n}$. Now, $\mathbb{P}(X = 0) \leq \frac{\mathrm{Var}(X)}{(\mathbb{E}[X])^2}$. So it suffices to show that $\frac{\mathrm{Var}(X)}{(\mathbb{E}[X])^2} \to 0$. We have

$$\begin{aligned} X &= \sum_{K\in\{0,\dots,n\}^{(3)}} \mathbb{1}(K \text{ is a triangle in } G) \\ X^2 &= \sum_{K\in\{0,\dots,n\}^{(3)}} \sum_{L\in\{0,\dots,n\}^{(3)}} \mathbb{1}(K, L \text{ are triangles in } G) \\ \mathbb{E}\left[X^2\right] &= \sum_{K\in\{0,\dots,n\}^{(3)}} \sum_{L\in\{0,\dots,n\}^{(3)}} \mathbb{P}(K, L \text{ are triangles in } G) \end{aligned}$$

and

$$(\mathbb{E}[X])^2 = \sum_{K\in\{0,\dots,n\}^{(3)}} \sum_{L\in\{0,\dots,n\}^{(3)}} \mathbb{P}(K \text{ is a triangle in } G)\,\mathbb{P}(L \text{ is a triangle in } G)$$

When computing $\mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$, the only terms that do not cancel are those terms which share edges.

$$\begin{aligned} \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 \leq & \sum_{K\in\{0,\dots,n\}^{(3)}} \sum_{L \text{ that shares a single edge with } K} \mathbb{P}(K, L \text{ are triangles in } G) \\ & + \sum_{K\in\{0,\dots,n\}^{(3)}} \mathbb{P}(K \text{ is a triangle in } G) \\ & \leq \sum_{K\in\{0,\dots,n\}^{(3)}} \sum_{L \text{ that shares a single edge with } K} \mathbb{P}(K, L \text{ are triangles in } G) + \mathbb{E}[X] \\ & \leq \underbrace{n^4 p^5}_{fourvertices, fiveedges} + \mathbb{E}[X] \end{aligned}$$

Hence,

$$\frac{\operatorname{Var}(X)}{(\mathbb{E}[X])^2} \leq \frac{n^4p^5 + \mathbb{E}[X]}{(\mathbb{E}[X])^2} \leq X\frac{n^4p^5}{(p^3n^3)^2} + \frac{1}{\mathbb{E}[X]} \leq \frac{1}{pn^2} + \frac{1}{\mathbb{E}[X]} \to 0$$

□

Remark. We see a 'phase transition' from in $\mathbb{P}(K_3 \subset G)$ as p moves from below $\frac{1}{n}$ to above $\frac{1}{n}$. Suppose $p = \frac{\lambda}{n}$ for some fixed $\lambda > 0$. Here, $\lim_{n\to\infty} \mathbb{P}(K_3 \subset G) = 1 - e^{-\frac{\lambda^3}{6}}$, but this result will not be proven.

Remark. We have seen that if the expected number of triangles increases to infinity, then the probability that $G \sim G(n,p)$ contains a triangle converges to 1. However, this is not true in general, replacing 'triangle' with another graph. Consider the graph H defined by a triangle with 1000 extra disjoint vertices. Here, the expected amount of copies of H is $\binom{n}{1003}p^3 \approx \frac{n^{1003}}{1003!}p^3$, which becomes large when $p = n^{-\frac{1003}{3}} < \frac{1}{n}$. If K is the 'densest' subgraph of H, then if the expected amount of copies of K tends to infinity, the probability that G contains a copy of H tends to 1.

6.4. Connectedness

Throughout this section, we will use the inequality $1 - x \leq e^{-x}$.

Proposition. Let $G \sim G(n,p)$. Then, for all $\varepsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}(G \text{ has an isolated vertex}) = \begin{cases} 0 & p \geq (1+\varepsilon)\frac{\log n}{n} \\ 1 & p \leq (1-\varepsilon)\frac{\log n}{n} \end{cases}$$

where a vertex is *isolated* if its degree is zero.

Proof. Let I be the number of isolated vertices in G. Then,

$$\mathbb{E}[I] = \sum_{i=1}^{n} \mathbb{P}(v_i \text{ is isolated}) = \sum_{i=1}^{n}(1-p)^{n-1} = n(1-p)^{n-1}$$

If $p \geq (1+\varepsilon)\frac{\log n}{n}$, then

$$\mathbb{E}[I] = \frac{n(1-p)^n}{1-p} \leq ne^{-pn} \leq ne^{-(1+\varepsilon)\frac{\log n}{n}n} = ne^{-(1+\varepsilon)\log n} = nn^{-(1+\varepsilon)} = n^{-\varepsilon} \to 0$$

Hence, by Markov's inequality, the probability that G has an isolated vertex is $\mathbb{P}(I \geq 1) \leq \mathbb{E}[I] \to 0$. If $p \leq (1-\varepsilon)\frac{\log n}{n}$, then

$$\mathbb{E}[I] = \frac{n(1-p)^n}{1-p} \geq n(1-p)^n \geq ne^{-\left(1+\frac{\varepsilon}{4}\right)pn}$$

for sufficiently large n, and sufficiently small ε. This statement holds because $1-p = e^{\log(1-p)}$ and Taylor's theorem implies $\log(1-p) = -p + \frac{p^2}{2} + o(p^2)$. Then

$$\mathbb{E}\left[I\right] \geq ne^{-\left(1+\frac{\varepsilon}{4}\right)(1-\varepsilon)\log n} = nn^{-\left(1+\frac{\varepsilon}{4}\right)(1-\varepsilon)} = nn^{-1+\frac{3\varepsilon}{4}+\frac{\varepsilon^2}{4}} = n^{\frac{3\varepsilon}{4}+\frac{\varepsilon^2}{4}} \to \infty$$

We will apply the second moment method on I. We have $\mathbb{P}\left(I = 0\right) \leq \frac{\mathrm{Var}(I)}{(\mathbb{E}[I])^2}$.

$$\begin{aligned}
\mathrm{Var}\left(I\right) &= \mathbb{E}\left[I^2\right] - \left(\mathbb{E}\left[I\right]\right)^2 \\
&= \sum_{u,v\in V(G)} \mathbb{P}\left(d(u) = 0, d(v) = 0\right) - \sum_{u,v\in V(G)} \mathbb{P}\left(d(u) = 0\right)\mathbb{P}\left(d(v) = 0\right) \\
&\leq \mathbb{E}\left[I\right] + \sum_{u\neq v} \left(\mathbb{P}\left(d(u) = 0, d(v) = 0\right) - \mathbb{P}\left(d(u) = 0\right)\mathbb{P}\left(d(v) = 0\right)\right) \\
&= \mathbb{E}\left[I\right] + \sum_{u\neq v} \left((1-p)^{2(n-1)} - (1-p)^{2(n-1)}\right) \\
&\leq \mathbb{E}\left[I\right] + n^2(1-p)^{2(n-1)}\left(\frac{1}{1-p} - 1\right) \\
\frac{\mathrm{Var}\left(I\right)}{\left(\mathbb{E}\left[I\right]\right)^2} &\leq \frac{1}{\mathbb{E}\left[I\right]}\frac{1}{n^2(1-p)^{2(n-1)}}n^2(1-p)^{2(n-1)}\left(\frac{1}{1-p} - 1\right) \\
&\leq \frac{1}{\mathbb{E}\left[I\right]} + \frac{1}{1-p} - 1 \to 0
\end{aligned}$$

since $p \to 0$ and $\mathbb{E}\left[I\right] \to \infty$ for $p < (1-\varepsilon)\frac{\log n}{n}$, as required. □

Theorem. Let $G \sim G(n,p)$. Then for all $\varepsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(G \text{ connected}\right) = \begin{cases} 1 & p \geq (1+\varepsilon)\frac{\log n}{n} \\ 0 & p \leq (1-\varepsilon)\frac{\log n}{n} \end{cases}$$

Remark. This is an example of a *sharp threshold*. Above, we saw the *coarse threshold* $p \gg \frac{1}{n}$ and $p \ll \frac{1}{n}$. Often, sharp thresholds are seen in relation to global properties, and coarse thresholds are seen when analysing local properties.

Proof. Suppose $p \leq (1-\varepsilon)\frac{\log n}{n}$. We want to show that $\lim_{n\to\infty} \mathbb{P}\left(G \text{ connected}\right)$ converges to zero. This follows from the fact that $\mathbb{P}\left(G \text{ connected}\right) \geq \mathbb{P}\left(G \text{ has no isolated vertex}\right) \to 0$.

Now suppose $p \geq (1+\varepsilon)\frac{\log n}{n}$. We now want to show that $\lim_{n\to\infty} \mathbb{P}\left(G \text{ connected}\right)$ converges to one. If G is not connected, we can find $A \subset V(G)$ where $1 \leq |A| \leq \frac{n}{2}$, and there are no edges between A and $V(G) \setminus A$. Consider

$$
\begin{aligned}
\mathbb{P}\left(G \text{ not connected}\right) &= \mathbb{P}\Big(\exists A \subset V(G),\, 0 < |A| \leq \frac{n}{2}, e(A, V(G) \setminus A) = 0\Big) \\
&= \mathbb{P}\left(\bigcup_{A \subset V(G), 0 < |A| \leq \frac{n}{2}} \{e(A, V(G) \setminus A) = 0\} \right) \\
&\leq \sum_{A \subset V(G),\, 0 < |A| \leq \frac{n}{2}} \mathbb{P}\left(e(A, V(G) \setminus A) = 0\right) \\
&= \sum_{A \subset V(G),\, 0 < |A| \leq \frac{n}{2}} (1-p)^{|A|(n-|A|)} \\
&= \sum_{k=1}^{\left\lfloor \frac{n}{2} \right\rfloor} \binom{n}{k} (1-p)^{k(n-k)} \\
&\leq \sum_{k=1}^{\frac{\varepsilon n}{4}} \binom{n}{k} (1-p)^{k(n-k)} + \sum_{k=\frac{\varepsilon n}{4}} \binom{n}{k} (1-p)^{k(n-k)} \\
&\leq \sum_{k=1}^{\frac{\varepsilon n}{4}} n^k e^{-pk(n-k)} + \sum_{k=\frac{\varepsilon n}{4}} \binom{n}{k} (1-p)^{k(n-k)} \\
&\leq \sum_{k=1}^{\frac{\varepsilon n}{4}} \left(n e^{-(1+\varepsilon)\frac{\log n}{n}(n-k)} \right)^k + \sum_{k=\frac{\varepsilon n}{4}} \binom{n}{k} (1-p)^{k(n-k)} \\
&\leq \sum_{k=1}^{\frac{\varepsilon n}{4}} \left(n e^{-(1+\varepsilon)\frac{\log n}{n} n\left(1-\frac{\varepsilon}{4}\right)} \right)^k + \sum_{k=\frac{\varepsilon n}{4}} \binom{n}{k} (1-p)^{k(n-k)} \\
&\leq \sum_{k=1}^{\frac{\varepsilon n}{4}} \left(n e^{-(1+\varepsilon)\log n\left(1-\frac{\varepsilon}{4}\right)} \right)^k + \sum_{k=\frac{\varepsilon n}{4}} \binom{n}{k} (1-p)^{k(n-k)} \\
&\leq \sum_{k=1}^{\frac{\varepsilon n}{4}} \left(n e^{-\log n\left(1+\frac{3\varepsilon}{4}\right)} \right)^k + \sum_{k=\frac{\varepsilon n}{4}} \binom{n}{k} (1-p)^{k(n-k)} \\
&\leq \sum_{k=1}^{\frac{\varepsilon n}{4}} \underbrace{\left(n^{-\left(1+\frac{3\varepsilon}{4}\right)} \right)^k}_{\to 0} + \sum_{k=\frac{\varepsilon n}{4}} \binom{n}{k} (1-p)^{k(n-k)} \\
&\leq \sum_{k=1}^{\frac{\varepsilon n}{4}} \left(n^{-\left(1+\frac{3\varepsilon}{4}\right)} \right)^k + \sum_{k=\frac{\varepsilon n}{4}} 2^n e^{-(1+\varepsilon)\frac{\log n}{n} k(n-k)} \\
&\leq \sum_{k=1}^{\frac{\varepsilon n}{4}} \left(n^{-\left(1+\frac{3\varepsilon}{4}\right)} \right)^k + \sum_{k=\frac{\varepsilon n}{4}} \underbrace{2^n e^{-(1+\varepsilon)\frac{\log n}{n}\frac{\varepsilon n}{4}\frac{n}{2}}}_{\to 0}
\end{aligned}
$$

as required. □

7. Algebraic graph theory

7.1. Graphs of a given diameter

Definition. Let G be a connected graph. The *diameter* of G is

$$\operatorname{diam} G = \max\{d(x,y) \mid x,y \in V(G)\}$$

Remark. The diameter of G is 1 if and only if G is complete, so there are $\binom{n}{2}$ edges.

Proposition. Let G be a graph with diameter at most 2. Then $|G| \leq \Delta(G)^2 + 1$.

Proof. Let $x \in G$. Then $V(G) = \{x\} \cup N(x) \cup N(N(x)) \setminus N(x)$. Hence $|G| \leq 1 + \Delta(G) + \Delta(G)(\Delta(G) - 1) \leq \Delta(G)^2 + 1$. □

Definition. A *Moore graph* is a graph for which $|G| = \Delta(G)^2 + 1$.

Remark. Any Moore graph is regular. Such a graph does not contain a triangle. A graph G is a Moore graph if and only if every distinct $x, y \in V(G)$ have a unique path of length at most 2 between them.

Example. C_5 is a Moore graph with $\Delta(C_5) = 2$. The Petersen graph is a Moore graph with degree 3.

7.2. Adjacency matrices

Definition. The *adjacency matrix* of a graph G on vertex set $\{1, \dots, n\}$ is the $n \times n$ matrix A_G with entries $a_{xy} = \mathbb{1}_{xy \in E(G)}$.

Remark. Adjacency matrices are symmetric and have zero diagonal, hence $\operatorname{tr} A_G = 0$.

Proposition. Let G be a graph, and A_G be its adjacency matrix. Let $k \in \mathbb{N}$. Then $(A_G^k)_{xy}$ is the number of walks of length k from x to y in G.

Proof. If $k = 1$, then the theorem clearly holds. If $k = 2$, then $(A_G^2)_{xy} = \sum_z (A_G)_{xz}(A_G)_{zy} = \sum_z \mathbb{1}_{x\sim z \in E}\mathbb{1}_{z \in y}$ counts the amount of walks of length 2. For $k > 2$, we can proceed by induction. □

A_G acts on $\mathbb{R}^n$ as it is a linear map.

Example. Consider the graph C_4 on vertex set $\{1, 2, 3, 4\}$. This has adjacency matrix

$$A_{C_4} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

Let $x = (1, 2, -2, 3)^\intercal$. Then $A_G x = (5, -1, 5, -1)^\intercal$. Note that $(A_G x)_y$ is the sum of x_z for $z \sim y$.

Proposition. Let A be an $n \times n$ symmetric matrix. Then A has real eigenvalues λ_i, and there exists an orthonormal basis u_i where $Au_i = \lambda_i u_i$.

Given a graph G on n vertices, we can now consider its eigenvalues and eigenvectors, which are the eigenvalues and eigenvectors of A_G. Let $\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n = \lambda_{\min}$ without loss of generality. Since $\sum_{i=1}^n \lambda_i = \operatorname{tr} A_G = 0$, if G is a nonempty graph, $\lambda_{\max} > 0$ and $\lambda_{\min} < 0$.

Example. $(1,1,1,1)^\intercal$ is an eigenvector of C_4 with eigenvalue 2. Note that the rank of A_G is 2, so there are two zero eigenvalues. Since the eigenvalues sum to zero, $\lambda_{\min} = -2$. One example of a corresponding eigenvector is $(1,-1,1,-1)^\intercal$.

Proposition. Let A be a symmetric $n \times n$ matrix. Then

$$\lambda_{\max} = \max_{x \in \mathbb{R}^n \setminus \{0\}} \left(\frac{\langle x, Ax \rangle}{\langle x, x \rangle} \right); \quad \lambda_{\min} = \min_{x \in \mathbb{R}^n \setminus \{0\}} \left(\frac{\langle x, Ax \rangle}{\langle x, x \rangle} \right)$$

Proposition. Let G be a graph.

(i) If λ is an eigenvalue, then $|\lambda| \leq \Delta(G)$.

(ii) If G is connected, then $\Delta(G)$ is an eigenvalue if and only if G is regular. In this case, $\mathbb{1} = (1, \dots, 1)$ is the corresponding eigenvector, and $\Delta(G)$ has multiplicity 1.

(iii) If G is connected, then $-\Delta(G)$ is an eigenvalue if and only if G is regular and bipartite.

(iv) $\lambda_{\max} \geq \delta(G)$.

Proof. Part (i). Let λ be an eigenvalue for G. Let $x = (x_1, \dots, x_n)$ be a corresponding eigenvector. Let x_i be the entry with largest absolute value. We may assume that $x_i = 1$. Then, $\lambda x = Ax$ gives

$$\lambda = \lambda x_i = (\lambda x)_i = (Ax)_i = \sum_{j \sim i} x_j \implies |\lambda| \leq \left| \sum_{j \sim i} x_j \right| \leq \Delta(G)$$

Part (ii). Suppose G is regular. Then observe that $\mathbb{1} = (1, \dots, 1)$ is an eigenvector of G with eigenvalue $\delta(G) = \Delta(G)$. Now suppose $\Delta(G)$ is an eigenvalue. Let $x = (x_1, \dots, x_n)$ be a corresponding eigenvector and let x_i be the entry with largest absolute value. Without loss of generality let $x_i = 1$. We have $\Delta(G) = \Delta(G) x_i = \sum_{j \sim i} x_j$, so $\deg i = \Delta(G)$, and if $j \sim i$, then $x_j = 1$. Proceeding inductively, since the graph is connected, all x_j are equal to 1, and all vertices have degree $\Delta(G)$. So $x = \mathbb{1}$ as required. Since this is the only possible eigenvector with eigenvalue $\Delta(G)$, and A_G is symmetric, the multiplicity of the eigenvalue $\Delta(G)$ is 1.

Part (iii). Suppose G is bipartite and regular. Let $V(G) = X \sqcup Y$, and consider the vector given by $x_i = 1$ if $i \in X$ and $x_i = -1$ if $i \in Y$. Then $Ax = -\Delta(G)x$ as required. Now suppose $-\Delta(G)$ is an eigenvalue. As before, let x be an eigenvector with $x_i = 1$ of maximal absolute value. We have $-\Delta(G) = -\Delta(G) x_i = \sum_{j \sim i} x_j$, hence $\deg i = \Delta(G)$, and if $j \sim i$, we have

$x_j = -1$. Since G is connected, we repeat the process to show that G is $\Delta(G)$-regular, and x_j is either $+1$ or -1 giving a natural bipartition of the graph.

Part (iv). Note that

$$\lambda_{\max} = \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\langle x, Ax \rangle}{\langle x, x \rangle}$$

Consider $x = \mathbb{1} = (1, \dots, 1)$. Then

$$\lambda_{\max} \geq \frac{\langle \mathbb{1}, A\mathbb{1} \rangle}{\langle \mathbb{1}, \mathbb{1} \rangle} = \frac{1}{n} \sum_{i=1}^{n} \deg(i) \geq \delta(G)$$

□

7.3. Strongly regular graphs

Definition. A graph G is (k, a, b)*-strongly regular* if

(i) G is k-regular;

(ii) for every pair of adjacent vertices $x \sim y$, they have exactly a common neighbours, so $|N(x) \cap N(y)| = a$;

(iii) for every pair of not equal and non-adjacent vertices $x \nsim y$, they have exactly b common neighbours, so $|N(x) \cap N(y)| = b$.

Example. C_4 is $(2, 0, 2)$-strongly regular. C_5 is $(2, 0, 1)$-strongly regular. Any Moore graph is $(\Delta(K), 0, 1)$-strongly regular.

Theorem (strongly regular graphs are rare)**.** Let G be a (k, a, b)-strongly regular graph on n vertices. Then,

$$\frac{1}{2}\left((n-1) \pm \frac{(n-1)(b-a) - 2k}{\sqrt{(a-b)^2 + 4(k-b)}}\right)$$

are integers.

Proof. Let A be the adjacency matrix of G. Then

$$(A^2)_{xy} = \begin{cases} a & x \sim y \\ b & x \neq y, x \nsim y \\ k & x = y \end{cases} \implies A^2 = aA + b(J - I - A) + kI$$

where J is the matrix with $J_{xy} = 1$ for all x, y. Hence, $A^2 + (b-a)A + (b-k)I - bJ = 0$. We know that k is an eigenvalue of A, and the corresponding eigenvector is $\mathbb{1}$. Since G is connected, k has multiplicity 1.

Let λ be an eigenvalue of A such that $\lambda \neq k$. Let x be the corresponding eigenvector. Applying the matrix equation to x, we obtain $\lambda^2 x + (b-a)\lambda x + (b-k)x = 0$ as $Jx = 0$, as x is orthogonal to $\mathbb{1}$. Then $\lambda^2 + (b-a)\lambda + (b-k)$ as $x \neq 0$. Hence,

$$\lambda = \frac{(a-b) \pm \sqrt{(a-b)^2 + 4(k-b)}}{2}$$

In particular, there are only three possible eigenvalues for A, which are k and the two possible solutions to the quadratic equation for λ. Let λ, μ be the solutions to the above equation. Let λ have multiplicity s and μ have multiplicity t. Then,

$$0 = \operatorname{tr} A = \sum_{i=1}^{n} \lambda_i = s\lambda + t\mu + k$$

We also have $s + t + 1 = n$, since there are n eigenvalues. Solving both equations simultaneously, we obtain the result as desired. □

Corollary. Let G be a Moore graph with $\Delta(G) = k$. Then $k \in \{2, 3, 7, 57\}$.

Proof. If G is a Moore graph, it is $(k, 0, 1)$-strongly regular on $k^2 + 1$ vertices. Then, one can check the condition in the previous theorem.

$$\frac{1}{2}\left(k^2 \pm \frac{k^2 - 2k}{\sqrt{4k-3}}\right) \in \mathbb{Z}$$

□

Remark. It is not known if such a graph G with $k = 57$ exists.

IV. Automata and Formal Languages

Lectured in Michaelmas 2022 by Prof. B. Löwe

Computation, or computability, is central to modern mathematics. However, we very rarely think about the precise definition of what it means for something to be 'computable'. There is an important difference between existence and algorithmic access to a witness. In this course, we discuss the precise definition of computability, and use it to prove that there is no algorithm to solve certain problems.

There are many possible ways to define computation and computability, and it is a remarkable fact that most 'reasonable' definitions of computable functions coincide. This is known as the Church–Turing thesis, and it allows us to reason about computation without being tied to a specific model, such as register machines, the Turing machine, or Church's recursive functions.

A language is a set of strings called words. Languages can be used to model countable sets, such as the set of powers of two, the set of primes, or the set of numbers which describe a register machine that determine if a given computation will halt or not. We will explore different types of language, and use computation theory to study which properties of languages can be determined algorithmically. We prove that there is a large class of languages for which there is an algorithm to determine if a given word lies in the language, but a much smaller class of languages for which we can determine algorithmically if they contain any words at all.

Contents

1. Introduction

1.1. Exposition

Computation, or computability, is central to modern mathematics. However, we very rarely think about the precise definition of what it means for something to be 'computable'. There is an important difference between existence and algorithmic access to a witness. Contrast the statements 'every polynomial of order n has a root', and 'there is an algorithm that, given a polynomial of order n, we can find a root'. In many cases, there is an existence proof but no algorithm to construct the relevant object.

In 1900, Hilbert gave a talk in Paris known as *Mathematical Problems*, in which he described a list of 100 problems to be worked on in the coming 100 years. One of these problems, the tenth, relates to an algorithm to determine whether solutions of Diophantine equations, those in $\mathbb{Z}[X]$, exist. In 1928, Ackermann wrote the book *Grundzüge der theoretischen Logik*, in which he described the famous *Entscheidungsproblem*: given a formula φ, determine whether φ is a tautology (true regardless of how the variables are interpreted).

In both cases, Hilbert expected that solutions to these questions exist. Positive solutions to such problems do not require a definition of words like 'algorithm' or 'procedure', because we can agree on what an algorithm is when we see an example. However, to disprove such statements, we need to rigorously define what an algorithm is, in order to rule all possible algorithms out.

1.2. Basic definitions

To talk about computation, we must first define the objects on which computation takes place. Naturally, one would assume the objects to be some kind of number, but even the above two examples do not have inputs as numbers; instead, we see polynomials and formulas. Modern computation relies on encodings of complicated objects as strings of a finite set of symbols, such as the bits 0 and 1. We use a similar approach, using a set Ω, which is usually assumed to be finite, called the set of *symbols*, and then we define $\Omega^\star$ to be the set of finite sequences of objects of Ω, called the set of Ω-*strings*.

1.3. Revisiting Numbers and Sets

Recall that a set X is called *countable* if there is a surjection $\mathbb{N} \to X$, and that X is called *infinite* if there is an injection $\mathbb{N} \to X$.

Proposition. If X is nonempty and countable, then $X^\star$ is infinite and countable.

Proof. Since $X \neq \varnothing$, there exists $x \in X$. $X^\star$ is infinite, as the function mapping $n \in \mathbb{N}$ to $\underbrace{xx \dots x}_{n \text{ times}}$ is injective. Because X is countable, there exists a surjection $\pi : \mathbb{N} \to X$. Each natural $k \in \mathbb{N}$ has a unique prime number decomposition $\prod_{i \in \mathbb{N}} p_i^{k_i}$ where $p_0 = 2, p_1 =$

$3, p_2 = 5, \dots$ are the primes indexed by the naturals. We will interpret the k_i as encoding a sequence of elements of X, taking care to preserve the relevance of zero. Reading k_0 as the length of a sequence, the sequence $(k_1, \dots, k_{k_0})$ is a sequence of naturals. We then obtain the sequence $(\pi(k_1), \dots, \pi(k_{k_0}))$ in $X^\star$. By surjectivity of π, the function we have constructed $k \mapsto (\pi(k_1), \dots, \pi(k_{k_0}))$ is also surjective. □

Theorem (Cantor's theorem)**.** Let X be infinite. Then its power set $\mathcal{P}(X)$ is uncountable.

Proof. A simple diagonalisation argument shows there is no surjection from the naturals to the power set $\mathcal{P}(X)$. □

Proposition. If X is countable, then the set $\mathrm{Fin}(X) \subseteq \mathcal{P}(X)$ of all finite subsets of X is countable.

Proof. We construct a surjection from $X^\star$ to $\mathrm{Fin}(X)$; then by composition with the surjection obtained in the first proposition we construct a surjection $\mathbb{N} \to \mathrm{Fin}(X)$. Consider the forgetful function $f : X^\star \to \mathrm{Fin}(X)$, mapping $(x_1, \dots, x_n)$ to $\{x_1, \dots, x_n\}$. Since X is countable, $\pi : \mathbb{N} \to X$ is surjective, hence for $x \in X$, $\pi^{-1}(x) \subseteq \mathbb{N}$ is a nonempty set of naturals. Therefore, let n_x be the least element of $\pi^{-1}(x)$. Then, given $F \in \mathrm{Fin}(X)$, consider the set $\{n_x \mid x \in F\}$, order it in the usual way, and represent this as a sequence. This is a sequence of naturals with $|F|$ elements, and its π-image is exactly F. □

1.4. Notation

We will use the following notational conventions.

- The natural numbers $\mathbb{N}$ are defined as $\{0, 1, 2, \dots\}$.
- We use the standard set-theoretic construction of naturals as Von Neumann ordinals, $n = \{0, 1, \dots, n-1\}$. Therefore, a natural is the set of all lower naturals.
- X^n is the set of sequence of X-strings of length n, defined as $X^n = n \to X$, treating n as a set as above.
- We write $|\alpha| = \mathrm{domain}(\alpha)$ for the length of a sequence.
- $X^0 = 0 \to X$ is a type with only one element ε, which is the empty sequence.
- We can write $X^\star = \bigcup_{n \in \mathbb{N}} X^n$.
- Truncation of a sequence $\alpha \in X^n$ to the length $k \leq n$ is exactly $\alpha|_k$: the unique sequence of length k such that $\alpha|_k \subseteq \alpha$.
- Concatenation of sequences $\alpha, \beta \in X^\star$ where $|\alpha| = m, |\beta| = n$, is denoted $\alpha\beta \in X^{m+n}$, defined piecewise in the natural way.
- By recursion, we define $\alpha^0 = \varepsilon$ and $\alpha^{n+1} = \alpha\alpha^n$.

- We identify the sequence of length one with its entry: $x \in X$ can represent the sequence $(x) \in X^1$.
- If $Y, Z \subseteq X^\star$, we write $YZ = \{\alpha\beta \mid \alpha \in Y, \beta \in Z\}$.
- Similarly, if $Y = \{\alpha\}$, we can write $\alpha Z = \{\alpha\beta \mid \beta \in Z\}$.
- If $f : X \to Y$, we can lift this function to the space $X^\star \to Y^\star$ functorially to the function $\hat{f}$. Often, the hat is omitted.

2. Rewrite systems

2.1. Definitions

Definition. Let Ω be a finite set of symbols, and let $\Omega^\star$ be the set of Ω-strings. We call elements of $\Omega^\star \times \Omega^\star$ *rewrite rules* or *production rules*. Such elements (α, β) are written $\alpha \to \beta$.

Informally, we interpret a rewrite rule $\alpha \to \beta$ as a procedure that replaces an occurrence of α in a string with β.

Definition. A pair $R = (\Omega, P)$ is called a *rewrite system* if P is a finite set of rewrite rules.

Proposition. If Ω is finite, there are only countably many rewrite systems on Ω.

Proof. $\Omega^\star$ is countable, so $\Omega^\star \times \Omega^\star$ is countable. Every P is an element of $\mathrm{Fin}(\Omega^\star \times \Omega^\star)$, hence this is countable. □

Definition. If $R = (\Omega, P)$ is a rewrite system, and $\sigma, \tau \in \Omega^\star$, we write $\sigma \overset{R}{\to}_1 \tau$, pronounced '$\sigma$ is rewritten to τ in one step' or 'R produces τ from σ in one step', if there exist $\alpha, \beta, \gamma, \delta \in \Omega^\star$ such that $\sigma = \alpha\gamma\beta$, $\tau = \alpha\delta\beta$, and $\gamma \to \delta \in P$.

The relation $\overset{R}{\to}$ is the reflexive and transitive closure of $\overset{R}{\to}_1$. The sequence $\sigma_0 \overset{R}{\to}_1 \sigma_1 \overset{R}{\to}_1 \dots \overset{R}{\to}_1 \sigma_n$ is called a *R-derivation of length n* of σ_n from σ_0. We write

$$\mathcal{D}(R, \sigma) = \left\{ \tau \in \Omega^\star \mid \sigma \overset{R}{\to} \tau \right\}$$

for the set of strings that can be rewritten, produced, or derived from σ.

2.2. Relation to languages

In language, we can think of Ω as representing letters, and $\Omega^\star$ representing words. We could alternatively consider Ω to represent words, and $\Omega^\star$ to represent sentences. Further, Ω could represent sentences, and then $\Omega^\star$ would represent texts.

However, not all elements of $\Omega^\star$ in each level is a valid word, sentence, or text. We therefore would like to describe which elements of $\Omega^\star$ are *well-formed*. Natural languages spoken by humans are finite, and normally the way we determine whether a string is a word is by consulting a dictionary, which at its core is a lookup table that determines whether any given string is or is not a word.

Even though in practice languages are finite, Chomsky realised that it makes more sense to model them as infinite sets, due to a property known as *linguistic recursion* that seems to be an important feature of human language. Linguistic recursion can be seen through the following example: when X is a sentence in English, 'E observes that X' is also a grammatical sentence in English. If we define an upper sentence length in English, we have to arbitrarily define an upper limit on this form of recursion.

There is a difference between a sentence being grammatical and being meaningful. One notable example is the grammatically correct ‘colourless green ideas sleep furiously’ that does not have meaning, to contrast with ‘furiously sleep ideas green colourless’ which is neither grammatically correct or meaningful. We can use grammar to distinguish these two sentences, but we cannot distinguish algebraically whether a sentence has meaning.

Example. Consider the following *generative grammar* of rewrite rules for English.

$$\begin{aligned} S &\to \text{NP VP} \\ \text{NP} &\to \text{Adj NP} \\ \text{NP} &\to \text{Noun} \\ \text{VP} &\to \text{Verb} \\ \text{VP} &\to \text{Verb Adv} \end{aligned}$$

This rewrite system allows us to derive the sentence ‘colourless green ideas sleep furiously’ from S.

2.3. Grammars

Definition. Let Σ be an *alphabet* of *letters* or *terminal symbols*, and let V be a set of *variables* or *nonterminal symbols*, such that Σ, V are nonempty and disjoint. Let $\Omega = \Sigma \cup V$. $a, b, c, \dots$ refer to letters and $A, B, C, \dots$ refer to variables. Elements of $\mathbb{W} = \Sigma^\star \subseteq \Omega^\star$ are called *words*. $u, v, w, \dots$ refer to words. We denote $\mathbb{W}^+ = \Sigma^\star \setminus \{\varepsilon\}$ for the set of nonempty words. A subset of $\mathbb{W}$ is called a *language*.

Note that there are uncountably many languages over any nonempty alphabet.

Definition. A tuple $G = (\Sigma, V, P, S)$ is called a *grammar* if Σ, V are nonempty and disjoint denoting $\Omega = \Sigma \cup V$, such that $R = (\Omega, P)$ is a rewrite system, and $S \in V$ is the *start symbol*. Since grammars give rise to a natural rewrite system, our notation for rewrite systems may also be used for grammars. For example,

$$\mathcal{D}(G, \sigma) = \mathcal{D}(R, \sigma); \quad \sigma \overset{G}{\to}_{(1)} \tau \iff \sigma \overset{R}{\to}_{(1)} \tau$$

We define the *language generated by the grammar* to be

$$\mathcal{L}(G) = \mathcal{D}(G, S) \cap \mathbb{W}$$

Example. If there is no rule of the form $S \to \alpha$ in P, then $\mathcal{D}(G, S) = \{S\}$ and thus $\mathcal{L}(G) = \varnothing$ because the start symbol is not a word. Likewise, if there is no rule of the form $\alpha \to w$ for $w \in \mathbb{W}$ in P, then $\mathcal{D}(G, S)$ contains no words, so $\mathcal{L}(G) = \varnothing$.

Example. Let $\Sigma = \{a\}$, $V = \{S\}$, $P_0 = \{S \to aaS, S \to a\}$, $G_0 = (\Sigma, V, P, S)$. We will show $\mathcal{L}(G_0) = \left\{a^{2n+1} \mid n \in \mathbb{N}\right\}$. First, every element of $\mathcal{D}(G, S)$ that is produced by G_0 is of odd length, which can be seen by induction on the length of the derivation, since each production rule preserves parity of length. Conversely, each a^{2n+1} can be produced by the rewrite rules, by applying $S \to aaS$ a total of n times, and then applying $S \to a$.

Note that the only requirement of the proof was that odd length is preserved. Thus, the following sets of production rules also produce the same language.

- $P_1 = \{S \to aSa, S \to a\}$
- $P_2 = \{S \to Saa, S \to a\}$
- $P_3 = \{S \to aaS, S \to aaSaa, S \to a\}$

This notion is called *equivalence of grammars*.

2.4. Equivalent grammars

Definition. Grammars G, G' are *equivalent* if $\mathcal{L}(G) = \mathcal{L}(G')$.

We intend to show that for a fixed finite set Σ, there are only countably many languages of the form $\mathcal{L}(G)$ for a grammar G (which may have arbitrary variable sets V).

Definition. Let $G = (\Sigma, V, P, S), G' = (\Sigma, V', P', S')$ be grammars on the same alphabet Σ. A function $f : \Omega \to \Omega' = \Sigma \cup V \to \Sigma \cup V'$ is called an *isomorphism* if

(i) $f|_\Sigma = \mathrm{id}$;

(ii) $f(S) = S'$;

(iii) $f|_V$ is a bijection from V to V';

(iv) $\alpha \to \beta \in P \iff f(\alpha) \to f(\beta) \in P'$.

Note that here, since $\alpha, \beta \in \Omega^\star$, $f(\alpha) = \hat{f}(\alpha)$ is the extension of f to $\Omega^\star$.

Proposition. Isomorphic grammars are equivalent.

Proof. If f is an isomorphism from G to G', f^{-1} is an isomorphism from G' to G. Thus, by antisymmetry of $\subseteq$, it suffices to show that $\mathcal{L}(G) \subseteq \mathcal{L}(G')$. Let $w \in \mathcal{L}(G)$. Then there is a derivation in G of w from S:

$$S = \sigma_0 \overset{G}{\longrightarrow}_1 \sigma_1 \overset{G}{\longrightarrow}_1 \dots \overset{G}{\longrightarrow}_1 \sigma_n = w$$

Applying f to each element of this sequence,

$$S' = f(\sigma_0) \overset{G'}{\longrightarrow}_1 f(\sigma_1) \overset{G'}{\longrightarrow}_1 \dots \overset{G'}{\longrightarrow}_1 f(\sigma_n) = w$$

The start and end symbols take these values due to property (i) and (ii). Each arrow holds by property (iv). This is a derivation of w from S' in G. Hence $w \in \mathcal{L}(G')$. □

Proposition. If $G = (\Sigma, V, P, S)$ and V' is such that $|V| = |V'|$, then there exist P', S' such that $\mathcal{L}(G) = \mathcal{L}(G')$ with $G' = (\Sigma, V', P', S')$.

Proof. Since $|V| = |V'|$, there exists a bijection $f : V \to V'$. Then, extending this to $\Omega = \Sigma \cup V$ by letting $f(a) = a$ for all $a \in \Sigma$, this satisfies properties (i) and (iii) of the definition of an isomorphism. Define $S' = f(S)$ and $P' = \{f(\alpha) \to f(\beta) \mid \alpha \to \beta \in P\}$, so that properties (ii) and (iv) are satisfied. Then (Σ, V, P, S) is isomorphic to (Σ, V', P', S') and thus they have the same language. □

Proposition. There are only countably many languages of the form $\mathcal{L}(G)$ for some grammar G on a fixed alphabet Σ.

Proof. Let $\mathcal{L}$ be the set of all such languages. For a fixed V, there are only countably many rewrite systems with this choice of Σ and V. Hence, the set $\mathcal{G}_V$ of all grammars with fixed V is a finite union (over all start symbols) of countable sets. Therefore $\mathcal{L}_V = \{\mathcal{L}(G) \mid G \in \mathcal{G}_V\}$ is also countable.

By the previous result, we can define $\mathcal{L}_n = \mathcal{L}_V$ for some n-element set V. Now, $\mathcal{L} = \bigcup_{n>0} \mathcal{L}_n$, which is a countable union of countable sets and is thus countable. □

Remark. The set of languages produced by grammars is countable, but the set of all languages $\mathcal{P}(\mathbb{W})$ is uncountable.

2.5. The Chomsky hierarchy

Production rules may have certain properties.

Definition. Let $\alpha \to \beta$ be a production rule. We call this rule:

(i) *noncontracting*, if $|\alpha| \leq |\beta|$;

(ii) *context-sensitive*, if $\exists A \in V$, $\exists \gamma, \delta \in \Omega^\star$, $\exists \eta \in \Omega^+$, $\alpha = \gamma A \delta, \beta = \gamma \eta \delta$;

(iii) *context-free*, if $\alpha = A \in V$ and $|\beta| > 0$;

(iv) *regular*, if $\alpha = A \in V$ and β is either $a \in \Sigma$ or $aB \in \Sigma V$.

Regular implies context-free, context-free implies context-sensitive, context-sensitive implies noncontracting. Let $\mathbb{Q}$ be any of the above four properties. We say that a grammar is $\mathbb{Q}$ if all its production rules are $\mathbb{Q}$. A language is $\mathbb{Q}$ if it admits a grammar which is $\mathbb{Q}$.

Theorem (Chomsky). A language is noncontracting if and only if it is context-sensitive.

Chomsky used the following notation: a language $\mathcal{L}$ is

- *type 0*, if it is of the form $\mathcal{L}(G)$ for some G;
- *type 1*, if it is of the form $\mathcal{L}(G)$ for some G context-sensitive;
- *type 2*, if it is of the form $\mathcal{L}(G)$ for some G context-free;
- *type 3*, if it is of the form $\mathcal{L}(G)$ for some G regular.

We can easily find production rules that are context-sensitive but not context-free, for example. However, it is less obvious to show that there is a *language* that can be defined using a context-sensitive grammar but no context-free grammar. One thing motivating our work will be the development of techniques to distinguish the different classes of languages in the Chomsky hierarchy.

2.6. Decision problems

We present three important decision problems.

(i) Consider the *word problem*. The input to this problem is a grammar and a word; the question is to determine whether the word lies in the language generated by the grammar.

(ii) The *emptiness problem* considers a grammar G. The question is whether $\mathcal{L}(G) = \varnothing$.

(iii) The *equivalence problem* asks whether two grammars G, G' are equivalent.

Definition. We call a problem *solvable* if there is an algorithm that gives the correct answer. Otherwise, we call such a problem *unsolvable*.

Posed for all grammars, all three problems above are unsolvable. However, when restricted to certain classes of the Chomsky hierarchy, the problems are more approachable.

Lemma. If G is a noncontracting grammar and $w \in \mathbb{W}$, there exists a bound $N \in \mathbb{N}$ depending only on $|w|$ and $|\Omega|$ such that $w \in \mathcal{L}(G)$ if and only if w has a G-derivation of length at most N.

Proof. Consider a G-derivation $S = \sigma_0, \dots, \sigma_n = w$ of w, and consider the length of each element of the sequence. As the grammar is noncontracting, the sequence $1 = |\sigma_0|, \dots, |\sigma_n| = |w|$ is nondecreasing. Consider a part of the derivation $\sigma_i, \dots, \sigma_{i+k}$ for which the length of the $|\sigma_i|$ does not change, so $|\sigma_i| = |\sigma_{i+k}|$. If $\sigma_r = \sigma_s$ for some $r \neq s \in \{i, \dots, i+k\}$, we can shrink the derivation to $\sigma_i, \dots, \sigma_r, \sigma_{s+1}, \dots, \sigma_{i+k}$.

Therefore, without loss of generality, we can assume $\sigma_0, \dots, \sigma_n$ is a derivation of minimal length, so all σ_i are distinct. Then by the pigeonhole principle,

$$n \leq \sum_{\ell=1}^{|w|} |\Omega|^{\ell} = N$$

□

Corollary. The word problem is solvable on noncontracting, context-sensitive, context-free, and regular grammars.

Proof. Let $w \in \mathbb{W}$. There is a finite, enumerable collection of possible derivations for w by the above lemma. Check each derivation manually. □

2.7. Closure problems

Closure problems are concerned with operations on languages to produce new languages. Let L, M be languages. Commonly used operations include $LM, L \cup M, L \cap M, L \setminus M, \mathbb{W}^+ \setminus L$. Note that we use $\mathbb{W}^+ \setminus L$ instead of L^c because noncontracting languages cannot contain the empty word. If $\mathcal{C}$ is a class of languages, such as the class of all regular languages, we say that $\mathcal{C}$ is *closed* under an operation if applying that operation to elements of $\mathcal{C}$ yields a result which also lies in $\mathcal{C}$. We would like to see which classes are closed under which operations. Note that some closure properties imply others; for instance, closure under complement and intersection implies closure under union by De Morgan's laws.

Definition. Let $G = (\Sigma, V, P, S), G' = (\Sigma, V', P', S')$ be grammars. Then $H = (\Sigma, V \cup V' \cup \{T\}, P^\star, T)$ is called the *concatenation grammar*, where T is a new variable, and

$$P^\star = P \cup P' \cup \{T \to SS'\}$$

$H' = (\Sigma, V \cup V' \cup \{T\}, P^{\star\star}, T)$ is called the *union grammar*, where T is a new variable, and

$$P^{\star\star} = P \cup P' \cup \{T \to S, T \to S'\}$$

Remark. $\mathcal{L}(G)\mathcal{L}(G') \subseteq \mathcal{L}(H)$ by construction, and $\mathcal{L}(G) \cup \mathcal{L}(G') \subseteq \mathcal{L}(H')$, but it is not true *a priori* that the converse holds, because P and P' could share some variables. We can assume that V, V' are disjoint by relabelling, but that is insufficient for the converses to hold, because there may be interaction on the level of letters in Σ, which cannot be relabelled. The concatenation grammar on context-free grammars is context-free, and the union grammar on regular languages is regular.

Definition. A production rule $\alpha \to \beta$ is *variable-based* if all symbols occurring in α are variables. A grammar is called variable-based if all its rules are variable-based.

Remark. Regular and context-free languages are variable-based. Context-sensitive languages are not all variable-based.

Lemma. Every grammar is equivalent to a variable-based grammar.

Proof. Let $G = (\Sigma, V, P, S)$. For each letter $a \in \Sigma$, we allocate a new variable X_a. We define the map $X : \Omega \to \Omega$ by $X(a) = X_a$ for $a \in \Sigma$, and $X(A) = A$ for $A \in V$. Then X extends in the natural way to a map $X : \Omega^\star \to \Omega^\star$. We can map each production rule in G to a version that uses only variables and no letters by applying X to both sides. Hence, we define $P' = \{X(\alpha) \to X(\beta) \mid \alpha \to \beta \in P\}$. Then, defining $P'' = \{X_a \to a \mid a \in \Sigma\}$, let $G' = (\Sigma, V \cup \{X_a \mid a \in \Sigma\}, P' \cup P'', S)$. This grammar is variable-based and so it suffices to show that it defines the same language as G.

Any G-derivation of w is transformed into a G'-derivation of $X(w)$ by the operation $\alpha \to X(\alpha)$. Similarly, if we have a G'-derivation that contains no letters anywhere, all strings occurring are of the form $X(\alpha)$ for some $\alpha \in \Omega^\star$, and the operation of replacing all occurrences of X_a with a transforms that derivation into a G-derivation. Thus, $w \in \mathcal{L}(G)$ if and only if $X(w) \in \mathcal{D}(G', S)$.

If $X(w) \in \mathcal{D}(G', S)$ then, by applying rules of the form $X_a \to a$ as needed, we have $w \in \mathcal{L}(G')$.

Conversely, suppose $w \in \mathcal{L}(G')$ and let $S = \sigma_0, \dots, \sigma_m = w$ be a G'-derivation of w. Applying the operation X to this derivation, we obtain a sequence $S = \tau_0, \dots, \tau_m$. This sequence is not necessarily a G'-derivation. If $\sigma_i \overset{G'}{\longrightarrow}_1 \sigma_{i+1}$ was an application of a rule of the form $X(\alpha) \to X(\beta)$, then the same rule gives $X(\sigma_i) \overset{G'}{\longrightarrow}_1 X(\sigma_{i+1})$. In the other case, $\sigma_i \overset{G'}{\longrightarrow}_1 \sigma_{i+1}$ was an application of a rule of the form $X_a \to a$, so applying X gives $X(\sigma_i) = X(\sigma_{i+1})$. Since for each letter a there is only one production rule that produces a, we know that $|w|$-many steps of the derivation must be of this form. Thus, removing these steps will make the remainder of the sequence $\tau_0, \dots, \tau_m$ a G'-derivation of length $m - |w|$ of $X(w)$. Then $w \in \mathcal{L}(G)$ as required. □

Remark. The classes of context-free, context-sensitive, and noncontracting grammars are stable under the action of turning a grammar into its equivalent variable-based grammar; all added rules are regular. Regularity is not necessarily preserved.

Theorem. Let $G = (\Sigma, V, P, S), G' = (\Sigma, V', P', S')$ be variable-based grammars with $V \cap V' = \varnothing$. Then $\mathcal{L}(H) = \mathcal{L}(G)\mathcal{L}(G')$, and $\mathcal{L}(H') = \mathcal{L}(G) \cup \mathcal{L}(G')$. The classes of regular, context-free, context-sensitive, and noncontracting languages are closed under union. The classes of context-free, context-sensitive, and noncontracting languages are closed under concatenation.

Proof. First, convert G, G' into variable-based grammars with disjoint variable sets. Then, the concatenation grammar and union grammar for G, G' must produce the required language by disjointness. □

2.8. The empty word

Remark. By induction, we can easily show that noncontracting grammars cannot produce the empty word, so we usually work with $\mathbb{W}^+$ in place of $\mathbb{W}$. In general, adding the rule $S \to \varepsilon$ to a grammar in order to allow the empty word may introduce side-effects due to reuse of S.

Definition. A rule $\alpha \to \beta$ is *S-safe* if it does not contain S in β. A grammar is *ε-adequate* if all rules are S-safe.

An ε-adequate grammar admits the addition of the rule $S \to \varepsilon$ converting the language $\mathcal{L}(G)$ into $\mathcal{L}(G) \cup \{\varepsilon\}$. We can easily convert a grammar into an equivalent ε-adequate grammar by mapping $G = (\Sigma, V, P, S)$ to $G' = (\Sigma, V \cup \{T\}, P \cup \{T \to S\}, T)$ where T is a new variable.

3. Regular languages

3.1. Regular derivations

Definition. A rule of the form $A \to a$ is called a *terminal rule*. A rule of the form $A \to aB$ is called a *nonterminal rule*.

Lemma. Let G be a regular grammar. If $S \xrightarrow{G} \alpha$, then $\alpha \in \mathbb{W} \cup \mathbb{W}V$.

Proof. This is shown by induction on the length of the derivation. The length-zero derivation gives $\alpha = S = \varepsilon S \in \mathbb{W}V$. Suppose $S \xrightarrow{G} \beta \xrightarrow{G}_1 \alpha$ where $\beta \in \mathbb{W} \cup \mathbb{W}V$. If $\beta \in \mathbb{W}$, β contains no variables, but all rules rewrite a variable. This contradicts that $\beta \xrightarrow{G}_1 \alpha$. So suppose $\beta = wA$ for $w \in \mathbb{W}, A \in V$. Then the rule must be of the form $A \to a$ or $A \to aB$. Hence $\beta = wa$ or $\beta = waB$. In either case, the required invariant holds. □

Lemma. If $S \xrightarrow{G} w$ for $w \in \mathbb{W}$, then the derivation has length $|w|$ and consists of precisely $|w| - 1$ nonterminal rules and one final terminal rule.

Proof. Terminal rules preserve the length of a string, and decrement the amount of variables. Nonterminal rules increment the length of a string, and preserve the amount of variables. Given that S is a string of length one with one variable, we must apply $|w| - 1$ nonterminal rules to increment the length of the string $|w| - 1$ times. By the previous lemma, the number of variables in each derived string is always 0 or 1. If the number ever reaches zero, nothing can be rewritten. Given $w \in \mathbb{W}$, the number must reach zero, so a single terminal rule must be applied at the end. □

Note that the derivation is not uniquely determined.

Lemma. Regular languages are closed under concatenation.

Proof. Let $G = (\Sigma, V, P, S), G' = (\Sigma, V', P', S')$, where without loss of generality $V \cap V' = \varnothing$. Let $P^\star$ be the set of production rules given by P, but for each terminal rule $A \to a$ in P, replace it with a nonterminal rule $A \to aS'$. Then let $H = (\Sigma, V \cup V', P^\star \cup P', S)$. We claim $\mathcal{L}(H) = \mathcal{L}(G)\mathcal{L}(G')$.

Suppose $S \xrightarrow{G} v$ and $S' \xrightarrow{G'} w$. Then $S \xrightarrow{H} vS'$, and so $S \xrightarrow{H} vw$ as required.

Conversely, suppose $S \xrightarrow{u}$ for $u \in \mathbb{W}$. By the above lemma, the derivation is of the form

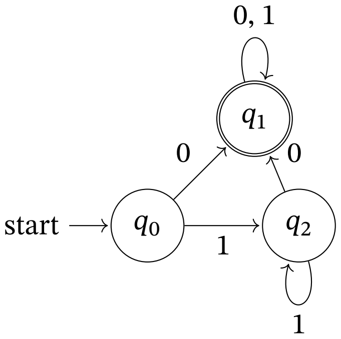
$$S = \sigma_0 \xrightarrow{H}_1 \dots \xrightarrow{H}_1 \sigma_n = u$$

where $\sigma_i = w_i X_i$ for some $w_i \in \mathbb{W}, X_i \in V$. An initial segment of the X_i belongs to V, until rewritten as S' by a rule added into $P^\star$. Then, the rest of the X_i belong to V', because only the new rules in $P^\star$ map variables between V and V'. Hence the derivation splits into two halves, $u = vw$ where $S \xrightarrow{G} v, S' \xrightarrow{G'} w$, giving the concatenation as required. □

3.2. Deterministic automata

Definition. Let Σ be an alphabet. Then a *deterministic automaton* is a tuple of the form $D = (\Sigma, Q, \delta, q_0, F)$ where Q is a finite set of *states*, $q_0 \in Q$ is the *start state*, $F \subseteq Q \setminus \{q_0\}$ is the *accept states*, and $\delta : Q \times \Sigma \to Q$ is the *transition function*.

We graphically represent deterministic automata using labelled directed graphs. The nodes are elements of Q, circled twice for accept states and circled once for other states. Each node has $|\Sigma|$-many outgoing arrows labelled with the corresponding letter.



We intuitively interpret a deterministic automaton as a machine that starts at q_0 and reads a word $w \in \mathbb{W}$ symbol-by-symbol, transitioning to a new state according to δ at each step. After all symbols in the word are parsed, we check whether the machine lies in an accept state or not. We say the automaton *accepts* w if the final state is an accept state; otherwise, it *rejects* w.

Definition. We define by recursion a function $\hat{\delta} : Q \times \mathbb{W} \to Q$ by $\hat{\delta}(q, \varepsilon) = q$ and $\hat{\delta}(q, aw) = \hat{\delta}(\delta(q, a), w)$. The *language accepted by D* is

$$\mathcal{L}(D) = \left\{w \mid \hat{\delta}(q_0, w) \in F\right\}$$

The sequence of states produced from q_0 and reading w is uniquely determined and of length $|w| + 1$, known as the *state sequence* of the computation.

We claim that in the example above, $\mathcal{L}(D) = \{w \mid w \text{ contains at least one } 0\}$. Note that $\hat{\delta}(q_0, w) = q_0$ if and only if $w = \varepsilon$. There are three transitions in the diagram for the letter 0, but all such 0-transitions lead to q_1 hence every string with a zero goes to q_1. All transitions from q_1 go back to q_1, so any string containing a zero must end at q_1. All other strings are of the form 1111 ... 1, which end at q_2.

Definition. Let $D = (\Sigma, Q, \delta, q_0, F), D' = (\Sigma, Q', \delta', q_0', F')$ be deterministic automata. Then a map $f : Q \to Q'$ is called a *homomorphism* from D to D' if

(i) for all $q \in Q$ and $a \in \Sigma$, we have $\delta'(f(q), a) = f(\delta(q, a))$;

(ii) $f(q_0) = q_0'$;

(iii) $q \in F$ if and only if $f(q) \in F'$.

In particular, if f is bijective, it has an inverse and is called an *isomorphism*. We can show by induction that $\hat{\delta}'(f(q), w) = f(\hat{\delta}(q, w))$. Note that if a homomorphism f is not surjective, the states not in its range are not *accessible* from q_0'. If f is not injective, so $f(p) = f(q)$ for $p \neq q$, then we have $f(\hat{\delta}(p, w)) = \hat{\delta}'(f(p), w) = \hat{\delta}'(f(q), w) = f(\hat{\delta}(q, w))$; we will say that such states p, q are *indistinguishable*. We will observe that failure to be bijective only affects inaccessible states or pairs of indistinguishable states.

Proposition. Let f be a homomorphism (not *a priori* an isomorphism) from D to D'. Then $\mathcal{L}(D) = \mathcal{L}(D')$.

Proof. Let $w \in \mathcal{L}(D)$, so $\hat{\delta}(q_0, w) \in F$. Applying f, $f(\hat{\delta}(q_0, w)) = \hat{\delta}'(f(q_0), w) = \hat{\delta}'(q_0', w) \in F'$ as required. All implications are bi-implications, so the converse holds. □

Theorem. Any language of the form $\mathcal{L}(D)$ for a deterministic automaton D is regular.

Proof. Let $D = (\Sigma, Q, \delta, q_0, F)$, and define a grammar $G = (\Sigma, V, P, S)$ by $V = Q, S = q_0$, and

$$P = \{p \to aq \mid \delta(p, a) = q\} \cup \{p \to a \mid \delta(p, a) \in F\}$$

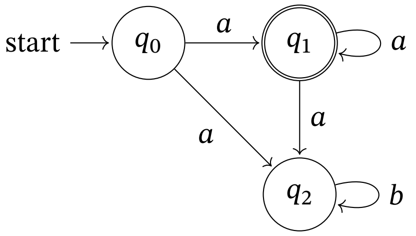
We will show $\mathcal{L}(D) = \mathcal{L}(G)$. Suppose $w = a_0 \dots a_n \in \mathcal{L}(D)$. Then $\hat{\delta}(q_0, w) \in F$, so there exist $q_0, \dots, q_{n+1}$ such that $q_{i+1} = \delta(q_i, a_i)$, and $q_{n+1} \in F$. By definition of G, this holds if and only if there exist $q_0, \dots, q_{n+1}$ such that $q_i \to a_i q_{i+1} \in P$ and $q_n \to a_n \in P$. This holds if and only if there exists $q_0, \dots, q_{n+1}$ such that $q_0 \overset{G}{\longrightarrow}_1 a_0 q_1 \overset{G}{\longrightarrow}_1 \dots \overset{G}{\longrightarrow}_1 a_0 \dots a_{n-1} q_n \overset{G}{\longrightarrow}_1 w$, so there exists a derivation $w \in \mathcal{L}(G)$. By regularity of G, all derivations are of this form, so we have bi-implications, and $\mathcal{L}(D) = \mathcal{L}(G)$. □

We will show that if L is a regular language, we can find a deterministic automaton D such that $L = \mathcal{L}(D)$. However, regular grammars can have multiple rules that may be used when reaching a single symbol, for instance $p \to aq$ and $p \to aq'$, so we cannot perform an obvious translation from this grammar into a deterministic automaton. To encapsulate this notion, we introduce nondeterministic automata.

3.3. Nondeterministic automata

Definition. A *nondeterministic automaton* is a tuple of the form $N = (\Sigma, Q, \delta, q_0, F)$ where Q is a finite set of states, $q_0 \in Q, F \subseteq Q \setminus \{q_0\}$, but in contrast with deterministic automata, we have $\delta : Q \times \Sigma \to \mathcal{P}(Q)$.

We interpret $\delta(q, a)$ as the set of possible states that the machine can transition into when reading a from state q. The graphical representation of such an automaton is the same.



Here, we define $\hat{\delta} : Q \times \mathbb{W} \to \mathcal{P}(Q)$, by the equations

$$\hat{\delta}(q, \varepsilon) = \{q\}; \quad \hat{\delta}(q, wa) = \bigcup_{p \in \hat{\delta}(q,w)} \delta(p, a)$$

This produces a unique *state set sequence*, not a deterministic state sequence. We define

$$\mathcal{L}(N) = \left\{w \mid \hat{\delta}(q_0, w) \cap F \neq \varnothing\right\}$$

Remark. Deterministic automata can be seen as a special case of nondeterministic automata.

Theorem. Let N be a nondeterministic automaton. Then there exists a deterministic automaton D such that $\mathcal{L}(N) = \mathcal{L}(D)$.

Our proof will involve a *subset construction*.

Proof. Let $N = (\Sigma, Q, \delta, q_0, F)$. We define $D = (\Sigma, \mathcal{P}(Q), \Delta, \{q_0\}, G)$, where

$$\Delta(X, a) = \bigcup_{q \in X} \delta(q, a); \quad G = \{X \in \mathcal{P}(Q) \mid X \cap F \neq \varnothing\}$$

We show that these two automata produce the same language. Consider the state sequence of D on input w.

$$X_0 = \{q_0\}; \quad X_{i+1} = \bigcup_{q \in X_i} \delta(q, a_i)$$

The state set sequence of N on input w is

$$Y_0 = \{q_0\}; \quad Y_{i+1} = \bigcup_{q \in Y_i} \delta(q, a_i)$$

Clearly, these exactly match, so $X_i = Y_i$. So w is accepted by D if and only if it is accepted by N. □

Remark. Although this construction always works, we have transformed an automaton on n states into one on 2^n states.

Theorem. Let G be a regular grammar. Then there exists a nondeterministic automaton N such that $\mathcal{L}(G) = \mathcal{L}(N)$.

Proof. Let $G = (\Sigma, V, P, S)$. Let $H \notin \Sigma \cup V$ be a new symbol, known as the *halt state*. Let $Q = V \cup \{H\}$. Define $N = (\Sigma, Q, \delta, S, \{H\})$ where

$$\delta(A, a) = \begin{cases} \{B \mid A \to aB \in P\} & \text{if } A \to a \notin P \\ \{B \mid A \to aB \in P\} \cup \{H\} & \text{if } A \to a \in P \end{cases}$$

We claim that $\mathcal{L}(G) = \mathcal{L}(N)$. If $w \in L(G)$, we have a sequence $A_0, \dots, A_{n+1}$ of variables such that

$$S = A_0 \overset{G}{\longrightarrow}_1 \dots \overset{G}{\longrightarrow}_1 a_0 \dots a_{n+1} A_{n+1} \overset{G}{\longrightarrow}_1 w$$

In particular, $A_i \to a_i A_{i+1} \in P$ and $A_{n+1} \to a_n \in P$. By definition of δ, there exists a sequence $A_1, \dots, A_{n+1}$ such that $A_{i+1} \in \delta(A_i, a_i)$ and $H \in \delta(A_n, a_n)$. Hence $H \in \hat{\delta}(S, w)$, so $w \in \mathcal{L}(N)$. All implications are bi-implications so the converse holds. □

3.4. The pumping lemma for regular languages

Definition. A language L *satisfies the regular pumping lemma with pumping number n* if every word $w \in L$ with length at least n can be split into three parts $w = xyz$, such that $|y| > 0$, $|xy| \leq n$ and for all $k \in \mathbb{N}$, we have $xy^kz \in L$. We call y a *pump* for the word xyz.

Theorem (regular pumping lemma)**.** Every regular language satisfies the pumping lemma.

Remark. If any word can be pumped, the language must be infinite.

Proof. Let L be a regular language. Then there exists a deterministic automaton D such that $L = \mathcal{L}(D)$. We show that L has pumping number $n = |Q|$. Let $w \in L(D)$ be a word with $|w| \geq n$. We can write $w = a_0 a_1 \dots a_{n-1} v$ where $v \in \mathbb{W}$.

The state sequence of D reading $a_0, \dots, a_{n-1}$ is $q_0, \dots, q_n$; it has length $n + 1$ since there are n state transitions. But there are only n states, so by the pigeonhole principle, one state must repeat. Let $i < j \leq n$ such that $q_i = q_j$. Let $x = a_0 \dots a_{i-1}, y = a_i \dots a_{j-1}, z = a_j \dots a_{n-1} v$, so we have $xyz = w$, $|y| > 0$, $|xy| \leq n$ by construction.

We show that we can pump the word. After reading x, we have $\hat{\delta}(q_0, x) = q_i$, and $\hat{\delta}(q_i, y) = q_j = q_i$, and finally $\hat{\delta}(q_i, z) = \hat{\delta}(q_j, z) \in F$. Hence, $\hat{\delta}(q_0, xy^k) = q_i$ by induction on k. In particular, $\hat{\delta}(q_0, xy^kz) \in F$ as required. □

Example. Let $L = \{0^k 1^k, k > 0\}$. We claim this is not a regular language. Suppose L is regular, and has pumping number N. Consider the word $0^N 1^N \in L$; this word has more than N letters, so the word can be pumped. The pump must lie in the first N letters, all of which are zeroes. Pumping down, $0^{N-\ell} 1^N \in L$ where ℓ is the length of the pump. This is a contradiction since the length of the pump is nonzero. Note that this language is context-free, so we know that the inclusion of regular languages in context-free languages is proper.

Example. Let $n > 0$, and let $L = \{0^n w, w \in \mathbb{W}\}$. We show this is regular, but any deterministic automaton D such that $L = \mathcal{L}(D)$ has more than n states. For regularity, we can simply write down a grammar.

$$P = \{S \to 0X_0, X_0 \to 0X_1, \dots, X_{n-2} \to 0X_{n-1}, X_{n-2} \to 0,$$
$$X_{n-1} \to 0, X_{n-1} \to 1, X_{n-1} \to 0X_{n-1}, X_{n-1} \to 1X_{n-1}\}$$

This has exactly $n + 1$ states. Suppose that an automaton with at most n states has the same language. Then L satisfies the pumping lemma with pumping number n. In particular, we can pump down the word 0^n, obtaining a word with fewer zeroes, and this is not in the language.

Example. Some non-regular languages also satisfy the pumping lemma. Let $\Sigma = \{0, 1\}$. If a word $w \in \mathbb{W}$ contains at least one zero, we say the *tail* of the word is the number of ones that follow the last zero. Let $X \subseteq \mathbb{N}$ be an arbitrary set of naturals, and define L_X to be the set of words that contain no zeroes, or have a tail which lies in X. If $X \neq Y$, we have $L_X \neq L_Y$, so L is an injection from $\mathcal{P}(\mathbb{N})$ to the space of languages on Σ. Since there are uncountably many $X \subseteq \mathbb{N}$, but there are only countably many regular languages, there must be some non-regular languages of the form L_X.

We claim that all L_X satisfy the pumping lemma, so then there must be some L_X which are non-regular which satisfy the pumping lemma. Let $X \subseteq \mathbb{N}$; we claim this has pumping number 2. Let $w \in L_x$ such that $|w| \geq 2$.

Suppose w starts with a zero, so $w = 0z$. Then let $x = \varepsilon, y = 0$, so $w = xyz$. Pumping up does not change the tail; pumping down either does not change the tail or there are now no zeroes, but in either case, the new word lies in the language.

Conversely, suppose w starts with a one, so $w = 1z$. Let $x = \varepsilon, y = 1$, so $w = xyz$ as before. If z contains no zeroes, after pumping y, there are still no zeroes, so the new word is in the language. If z contains a zero, there is a tail, and pumping y does not influence the tail. Hence, the pumping lemma is satisfied.

3.5. Closure properties

We have already shown that regular languages are closed under concatenation and union. We will now show that they are closed under complement, intersection, and difference. For this, it suffices to show they are closed under complement, because intersection and difference can be expressed in terms of complement and union.

Consider an automaton $D = (\Sigma, Q, \delta, q_0, F)$. Without loss of generality, we can ensure that $\delta(q_i, a) \neq q_0$ for all i, a. Now define $D' = (\Sigma, Q, \delta, q_0, \mathbb{Q} \setminus (F \cup \{q_0\}))$. Then, $\mathcal{L}(D') = \mathbb{W}^+ \setminus \mathcal{L}(D)$.

There is an alternative construction to obtain union and intersection, known as the *product automaton* construction. Let $D = (\Sigma, Q, \delta, q_0, F)$ and $D' = (\Sigma, Q', \delta', q_0', F')$. We can define the pointwise product $D'' = (\Sigma, Q \times Q', \delta \times \delta', (q_0, q_0'), F'')$, where $(\delta \times \delta')((q, q'), a) =$

$(\delta(q,a), \delta'(q',a))$, and either $F'' = \{(q,q') \mid q \in F, q' \in F'\}$ or $F'' = \{(q,q') \mid q \in F \text{ or } q' \in F'\}$. We can see that the language generated by this new automaton is $\mathcal{L}(D) \cap \mathcal{L}(D')$ or $\mathcal{L}(D) \cup \mathcal{L}(D')$.

3.6. Emptiness problem

Lemma. Let L be a nonempty regular language with pumping number n. Then there is a word $w \in L$ such that $|w| < n$.

Proof. Let w be a word in L. If $|w| < n$, we are already done. Otherwise, it can be pumped down into a smaller word. By induction, we can obtain a word of length less than n. □

Corollary. The emptiness problem for regular grammars is solvable.

Proof. Given a regular grammar, we can obtain its pumping number. We can check every word below this length because the word problem is solvable; if no words are accepted, the language is empty. □

3.7. Regular expressions

Definition. The *Kleene star* operation on a language L, written $L^\star$, is given by

$$L^\star = \{w \mid \exists \text{ sequence of words in } L, w = \text{their concatenation}\}$$

In particular $\varepsilon \in L^\star$. The *Kleene plus* operation is $L^+ = L^\star \setminus \{\varepsilon\}$.

Definition. A *regular expression* on an alphabet Σ is defined inductively by:

(i) the symbol $\varnothing$ is a regular expression;

(ii) ε is a regular expression;

(iii) for all a in Σ, a is a regular expression;

(iv) if R, S are regular expressions, $(R + S)$ is a regular expression;

(v) if R, S are regular expressions, (RS) is a regular expression;

(vi) if R is a regular expression, $R^\star$ is a regular expression;

(vii) if R is a regular expression, R^+ is a regular expression.

By definition, nothing else is a regular expression. By recursion, we can assign a language $\mathcal{L}(E)$ to each regular expression E.

(i) $\mathcal{L}(\varnothing) = \varnothing$;

(ii) $\mathcal{L}(\varepsilon) = \{\varepsilon\}$;

(iii) for $a \in \Sigma, \mathcal{L}(a) = \{a\}$;

(iv) if R, S are regular expressions, $\mathcal{L}(R + S) = \mathcal{L}(R) \cup \mathcal{L}(S)$;

(v) if R, S are regular expressions, $\mathcal{L}(RS) = \mathcal{L}(R)\mathcal{L}(S)$;

(vi) if R is a regular expression, $\mathcal{L}(R^\star) = \mathcal{L}(R)^\star$;

(vii) if R is a regular expression, $\mathcal{L}(R^+) = \mathcal{L}(R)^+$.

Note that rules (iv) and (v) introduce parentheses, occasionally unnecessarily. When the meaning is unambiguous, these parentheses are omitted. The binding power of concatenation RS is higher than union $R + S$, so we can write $RS + T$ for $((RS) + T)$.

We say that a language is *essentially regular* if there is a regular language L' such that $L = L'$ or $L = L' \cup \{\varepsilon\}$.

Theorem. If E is a regular expression, $\mathcal{L}(E)$ is essentially regular.

This is an equivalence, but the converse (often called Kleene's algorithm) is not required for this course.

Proof. Observe that (i), (ii), (iii) are essentially regular languages, so it suffices to show that essentially regular languages are closed under (iv), (v), (vi), (vii). We have already shown that regular languages are closed under union and concatenation, and the proof for essentially regular languages follows easily. Note that $\mathcal{L}(E^\star) = \mathcal{L}(E + E^+)$, so it suffices to show closure of regular languages under the Kleene plus; we can then perform case analysis to prove the same for essentially regular languages.

Let $G = (\Sigma, V, P, S)$ be a regular grammar. Let $P^+ = P \cup \{A \to aS \mid A \to a \in P\}$. It suffices to show that $G^+ = (\Sigma, V, P^+, S)$ has the language $\mathcal{L}(G^+) = \mathcal{L}(G)^+$.

Suppose $w \in \mathcal{L}(G)^+$, so $w = w_0 \dots w_n$ for $w_i \in \mathcal{L}(G)$. If $n = 0$, $w \in \mathcal{L}(G)$ and any derivation can be translated easily into G^+. Otherwise, suppose $w_0 \dots w_{n-1} \in \mathcal{L}(G^+)$ by induction. Therefore there is a derivation $S \xrightarrow{G^+} w_0 \dots w_{n-1}$. This derivation ends with a terminal rule $A \to a$, so we can replace it with a nonterminal rule $A \to aS$, giving $S \xrightarrow{G^+} w_0 \dots w_{n-1}S \xrightarrow{G} w_0 \dots w_{n-1}w_n$, so $w \in \mathcal{L}(G)$ as required.

Now suppose $w \in \mathcal{L}(G^+)$. Without loss of generality we can assume that G is ε-adequate, so S does not occur on the right-hand side of a rule. Suppose we have a derivation $S \xrightarrow{G^+} w$. Let n be the number of times that S occurs in the derivation. We then prove $w \in \mathcal{L}(G)^+$ by induction on n. n cannot be zero. Suppose all words $v \in \mathcal{L}(G^+)$ lie in $\mathcal{L}(G)^+$ if they have a derivation with $n - 1$ occurrences of S. Since $n \geq 1$, we have $S \xrightarrow{G^+} vS \xrightarrow{G^+} w$ where vS is the last occurrence of S in the derivation of w. In particular, $S \xrightarrow{G^+} v$ with $n-1$ occurrences, since the last rule of $S \xrightarrow{G^+} vS$ is one of the added rules in P^+. By induction, $v \in \mathcal{L}(G)^+$. Since $vS \xrightarrow{G^+} w$, we know that $w = vw'$ by considering the possible derivations in regular languages. Hence $S \xrightarrow{G^+} w'$ with only one occurrence of S at the start. In particular none of our new rules were used, so $S \xrightarrow{G} w'$, so $w' \in \mathcal{L}(G)^+$, hence $w \in \mathcal{L}(G)^+$. □

3.8. Minimisation of deterministic automata

Definition. A state q is called *accessible* if there is a word w such that $q = \hat{\delta}(q_0, w)$. A state that is not accessible is called *inaccessible*.

Definition. States q and q' are *distinguished* by a word w if $\hat{\delta}(q, w) \in F$ and $\hat{\delta}(q', w) \notin F$, or vice versa. States that are distinguished by some word are called *distinguishable*. States that are not distinguished by any word are called *indistinguishable*.

If $f : Q \to Q'$ is a homomorphism, then

(i) if p, q are distinguishable, $f(p) \neq f(q)$;

(ii) if $q' \in Q'$ is accessible, q' lies in the range of f.

In particular, if f is a homomorphism from D to D' and all pairs of nonequal states in D are distinguishable, f is injective; if all states in D' are accessible, f is surjective.

Definition. An automaton D is called *irreducible* if all pairs of nonequal states are distinguishable and all states are accessible.

Hence, any homomorphism between irreducible automata is an isomorphism.

Defining $q \sim q'$ if q and q' are indistinguishable, $\sim$ is an equivalence relation. As usual, we write $[q]$ for the equivalence class of states indistinguishable from q. We can therefore define the *quotient automaton* by

$$D/{\sim} = \left(\Sigma, Q/{\sim}, [\delta], [q_0], [F]\right); \quad [\delta]([q], a) = [\delta(q, a)]; \quad [F] = \{[q] \mid q \in F\}$$

Note that if an equivalence class contains an accept state, the class is completely contained in F, so being an accept state is a class property. The map $[\delta]$ is well-defined: indeed, if $q \sim q'$, we have $\delta(q, a) \sim \delta(q', a)$, because if $\delta(q, a) \nsim \delta(q', a)$, there would exist a word w that distinguishes these two states, but then aw would distinguish q and q'.

If $q \nsim q'$, we can show the two states are distinguished in the quotient automaton. By induction, $[\hat{\delta}]([q], w) = [\hat{\delta}(q, w)]$. Suppose without loss of generality that $\hat{\delta}(q, w) \in F$, $\hat{\delta}(q', w) \notin F$. Then $[\hat{\delta}]([q], w) \in [F]$, but $[\hat{\delta}]([q'], w) \notin [F]$. So w distinguishes $[q]$ and $[q']$. In particular, each pair of nonequal states is distinguishable.

Note further that $\mathcal{L}(D) = \mathcal{L}\left(D/{\sim}\right)$, because the quotient map $q \mapsto [q]$ is a homomorphism. If D had no inaccessible states, $D/{\sim}$ also has no inaccessible states, since the quotient map is surjective.

Theorem. For every deterministic automaton, there is an irreducible deterministic automaton I such that $\mathcal{L}(D) = \mathcal{L}(I)$.

Proof. Let $A \subseteq Q$ be the set of accessible states in D. Let $D^\star = \left(\Sigma, A, \delta|_{A\times\Sigma}, q_0, F \cap A\right)$. The inclusion map from $D^\star$ to D is a homomorphism, so their languages are the same. Now let $I = {}^{D^\star}\!/\!\sim$. By the above discussion, I is irreducible and has the same language as $D^\star$. □

Remark. The number of states in I is at most the number of states in D.

Theorem. If I, I' are irreducible deterministic automata and $\mathcal{L}(I) = \mathcal{L}(I')$, then I and I' are isomorphic.

Proof. It suffices to construct a homomorphism between the two automata, since any homomorphism between irreducible automata is an isomorphism. Let $I = (\Sigma, Q, \delta, q_0, F)$ and $I' = (\Sigma, Q', \delta', q_0', F')$, and without loss of generality let $Q \cap Q' = \varnothing$. We can extend $\sim$ to $Q \cup Q'$, by defining $q \sim q'$ if for all w, $\hat\delta(q, w) \in F$ if and only if $\hat\delta'(q', w) \in F'$. We know $q_0 \sim q_0'$, because by assumption, the languages of the two automata are the same.

We show that for all $q \in Q$, there exists $q' \in Q'$ such that $q \sim q'$. Let $\mathrm{sp}(q)$ be the length of the shortest path from q_0 to q. Since I is irreducible, this is well-defined and finite for all $q \in Q$. We prove this claim by induction on $\mathrm{sp}(q)$. The base case is $\mathrm{sp}(q) = 0$ so $q = q_0$, and we have already shown $q_0 \sim q_0'$ as required.

Now suppose $\mathrm{sp}(q) = k + 1$. Then there exists $p \in Q$ and $a \in \Sigma$ such that $\delta(p, a) = q$ and $\mathrm{sp}(p) = k$. By the induction hypothesis, we can find $p' \in Q'$ such that $p \sim p'$. Then let $q' = \delta'(p', a)$, then $q' \sim \delta(p, a) = q$. Hence each $q \in Q$ has a $q' \in Q'$ such that $q \sim q'$.

We now will show that if $q' \sim q \sim p'$, we have $q' = p'$. By transitivity, $q' \sim p'$, but by irreducibility of I', $q' = p'$.

Because of the above results, we can construct a function $f : Q \to Q'$ defined by $q \mapsto q'$ where $q \sim q'$. This is well-defined and unique. We now claim f is a homomorphism. Since $q_0 \sim q_0'$, we have $f(q_0) = q_0'$. The requirement $q \in F \iff f(q) \in F'$ follows by definition of $\sim$. Now fix $q \in Q$ and $q' = f(q)$, so $q \sim q'$. Then, $\delta(q, a) \sim \delta'(q', a)$, so $f(\delta(q, a)) \sim \delta'(q', a) = \delta'(f(q), a)$. □

Remark. There is a unique (up to isomorphism) irreducible automaton that accepts a given regular language, and its size is smaller than all other automata that accept the same language.

3.9. Equivalence problem

We have already solved the word problem for noncontracting grammars and the emptiness problem for regular grammars. To solve the equivalence problem, we will construct minimal automata for two given regular grammars, and check whether they are isomorphic; if so, the languages are the same, and otherwise, the languages are different. We must check that this idea can be formulated into an algorithm which must complete in finitely many steps.

Proposition. Let D be a deterministic automaton and $q \in Q$ a state. Then it is solvable whether q is accessible.

Proof. If there is a word w such that $\hat{\delta}(q_0, w) = q$, then the shortest such word has length at most $|Q|$, which can be easily proven using the technique from the pumping lemma. We can explicitly check each word of length at most $|Q|$. □

Theorem (the table filling algorithm)**.** Let D be a deterministic automaton and $q, q' \in Q$ states. Then the proposition $q \sim q'$ is solvable.

Proof. Form a matrix A with entries indexed by $Q \times Q$. The entry indexed by (q, q') contains information about distinguishability of q, q'. In particular, $A_{q,q'}$ contains either nothing or a word w distinguishing q and q'. Since $\sim$ is an equivalence relation, it suffices to consider the upper triangular part of the matrix, excluding the diagonal. To initialise the matrix, if $q \in F$ and $q' \notin F$ we set $A_{q,q'} = \varepsilon$, since the empty word distinguishes q, q'.

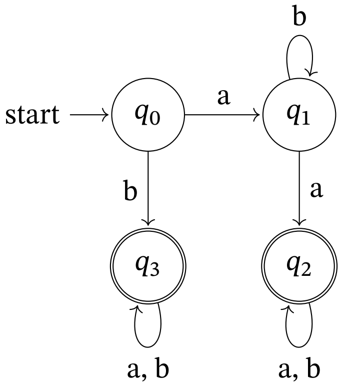
Then, for each $q, q' \in Q$ that do not have a filled entry $A_{q,q'}$ already, and for each $a \in \Sigma$, we can check the entry indexed by $(\delta(q, a), \delta(q', a))$. If these two states are distinguished by a word w, q and q' are distinguished by aw. So we can set $A_{q,q'} = aw$. This single step will terminate in a finite amount of time, on the order of $|Q|^2|\Sigma|$-many steps.

We then repeat this inductive step until no more assignments into the matrix can be made in an single iteration. This will happen in finitely many steps.

We now must show that after this process completes, $A_{q,q'}$ contains a word w if and only if q and q' are distinguishable, and in this case, w distinguishes q and q'. If $A_{q,q'}$ contains a word w, it is clear that w distinguishes q and q', since $\hat{\delta}(q, w) \in F$ and $\hat{\delta}(q', w) \notin F$ or vice versa. Now suppose there exists a word w that distinguishes some states q and q', but q, q' are unmarked in A. Let q, q' be a pair of states with a distinguishing word w of minimal length.

Either $w = \varepsilon$ or $w = av$. If $w = \varepsilon$, $q \in F$ and $q' \notin F$ or vice versa, so $A_{q,q'}$ is marked. Otherwise, $w = av$. Since v is shorter than the smallest word that distinguishes two states that are not marked in A, we must have that the entry $(\delta(q, a), \delta(q', a))$ is marked with some word in A. So at some step in the algorithm, this entry was added into A. But then the algorithm would mark q, q' with a word in the next step. □

Example. Consider the following automaton.



In step zero, we find

	q_0	q_1	q_2	q_3
q_0			ε	ε
q_1			ε	ε
q_2				
q_3				

In step one, checking $\delta(q_0, a)$ and $\delta(q_1, a)$, we arrive at

	q_0	q_1	q_2	q_3
q_0		a	ε	ε
q_1			ε	ε
q_2				
q_3				

The only remaining entry is (q_2, q_3), and this is not filled in a single step. Hence $q_2 \sim q_3$.

Corollary. The equivalence problem for regular grammars is solvable.

Hence, for regular grammars, all of our desirable closure properties are true, and all of our motivating decision problems are solvable.

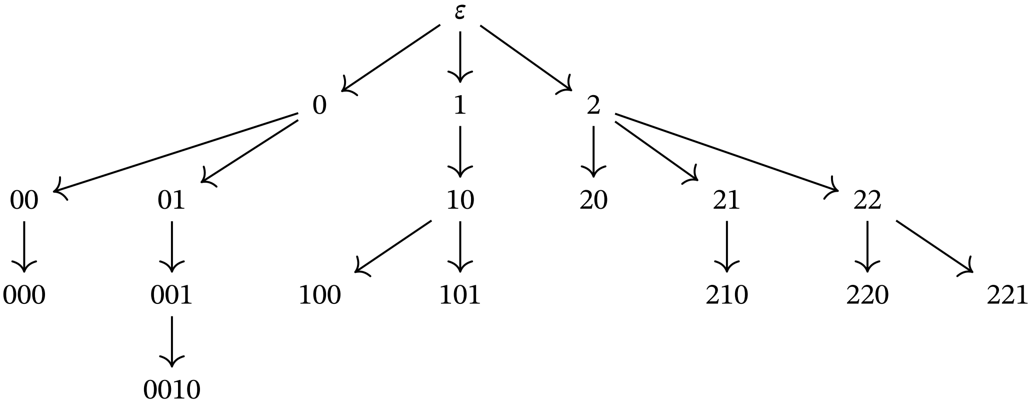
4. Context-free languages

4.1. Trees

Recall that the language $\{0^k 1^k \mid k > 0\}$ is context-free but not regular, so context-free languages are indeed a proper superset of regular languages. The structure of regular derivations was very simple; each intermediate step was of the form wA for a word w and a variable $A \in V$. However, the structure of context-free derivations is more complicated: we use a parse tree instead of a linear derivation.

Definition. A set $T \subseteq \mathbb{N}^*$ is called a *(finitely-branching) tree* if it is closed under initial segments, and for every $t \in T$, there is a *branching number* $n \in \mathbb{N}$ such that for all k, the sequence tk lies in T if and only if $k < n$. A node $t \in T$ with no sucessors is called a *leaf.* The empty sequence, which is an element of every tree, is called the *root.* A node $t \in T$ has *level* k if the length of the sequence is k, so $|t| = k$. If T is finite, there is a maximum level, called the *height* of the tree. For a node $t \in T$, the sequence $t|_0, t|_1, \dots, t|_{|t|} = t$ is called the *branch* leading to t.

Example. This is an example of a tree.



Definition. Let T be a tree and $t \in T$. Then $T_t = \{s \mid ts \in T\}$ is the *subtree starting from* t.

Definition. We define a partial order on T by $t < s$ if $t \neq s$ and if there exists k such that $t(k) \neq s(k)$ and k_0 is minimal with this property, then $t(k_0) < s(k_0)$. This is called the *left-to-right order.*

Remark. This order is only a partial order since it does not order two distinct nodes that lie on the same branch, for example, 0 and 00. For each level k, the nodes of length k are totally ordered. The leaves are totally ordered.

4.2. Parse trees

Definition. Let G be a context-free grammar. A pair $\mathbb{T} = (T, \ell)$ is a *G-parse tree* if T is a finitely-branching tree and $\ell : T \to \Omega$ is a *labelling function* such that:

(i) $\ell(\varepsilon) \in V$, we say T *starts with* $\ell(\varepsilon)$;

(ii) if $\ell(t) \in \Sigma$, t has no successors;

(iii) if t has $n+1$ successors and $\ell(t) = A \in V$, then $A \to \ell(t0)\ell(t1) \dots \ell(tn) \in P$.

If $\mathbb{T} = (T, \ell)$ is a G-parse tree, and $t_0, \dots, t_m$ are its leaves written in the left-to-right order, then the *string parsed by* $\mathbb{T}$ is $\sigma_{\mathbb{T}} = \ell(t_0) \dots \ell(t_m)$.

Remark. If $t \in T$, $\sigma_{\mathbb{T}} = \alpha\sigma_{\mathbb{T}_t}\beta$ where $\mathbb{T}_t = (T_t, \ell_t)$, $\ell_t(s) = \ell(ts)$.

Proposition. Let G be a context-free grammar. Then $w \in \mathcal{L}(G)$ if and only if there is a G-parse tree $\mathbb{T}$ starting from S such that $\sigma_{\mathbb{T}} = w$.

Proof. Observe that certain sequences of parse trees correspond to derivations. In particular, a sequence $\mathbb{T}_0, \dots, \mathbb{T}_n$ of G-parse trees is *derivative* if $\mathbb{T}_0 = (\{\varepsilon\}, \ell_0)$ with $\ell_0(\varepsilon) = S$, and $T_{i+1} \supseteq T_i$ is constructed by considering a leaf $t \in T_i$ such that $\ell_i(t) = A \in V$ and $A \to x_0 \dots x_n \in P$, and giving it $n+1$ successors with $\ell_{i+1}(tk) = x_k$. There is a one-to-one correspondence between G-derivations starting from S and such derivative sequences of parse trees. In particular, any derivation yields a derivative sequence of parse trees, and hence the last parse tree in the sequence has $\sigma_{\mathbb{T}_n} = w$.

Conversely, given a parse tree $\mathbb{T}$, it suffices to construct such a derivative sequence of parse trees, because then the correspondence yields a derivation as required. We start with the trivial tree $\mathbb{T}_0 = \left(\{\varepsilon\}, \ell|_{\{\varepsilon\}}\right)$. In each step, suppose $\mathbb{T}_0, \dots, \mathbb{T}_i$ already form a derivative sequence, and $T_i \neq T$. Let $t \in T \setminus T_i$. Then there is a terminal node in T_i on the branch containing t in T, which is not a terminal node in T. We can then create T_{i+1} by adding the T-successors of t to T_i. Since T is finite, after finitely many steps we are done. In particular, $\mathbb{T}_0, \dots, \mathbb{T}_n$ is a derivative sequence, and thus $S \overset{G}{\to} \sigma_{\mathbb{T}_n} = \sigma_{\mathbb{T}} = w$ as required. □

Suppose $\mathbb{T}$ is a parse tree and $t \in T$ such that $\ell(t) = A$, and $\mathbb{T}'$ is a parse tree starting from A. Then, we can remove the subtree T_t, and replace it with T', which also yields a parse tree. This technique is known as *grafting*.

Definition. We define $\operatorname{graft}(\mathbb{T}, t, \mathbb{T}') = (S, \ell^\star)$ where

$$S = \{s \in T \mid t \nsubseteq s\} \cup \{tu \mid u \in T'\}$$

and

$$\ell^\star(s) = \begin{cases} \ell(s) & t \nsubseteq s \\ \ell'(u) & \exists u \in T', s = tu \end{cases}$$

Then we have

$$\sigma_{\operatorname{graft}(\mathbb{T},t,\mathbb{T}')} = \alpha\sigma_{\mathbb{T}'}\beta; \quad \sigma_{\mathbb{T}} = \alpha\sigma_{\mathbb{T}_t}\beta$$

4.3. Chomsky normal form

Definition. A grammar is in *Chomsky normal form* if all of its rules are of the form $A \to BC$ or $A \to a$. Rules of the form $A \to BC$ are called *binary*; rules of the form $A \to a$ are called *unary*.

Every grammar in Chomsky normal form is context-free.

Lemma. Let G be a grammar in Chomsky normal form, and $w \in \mathcal{L}(G)$ with $|w| = n$. Then every G-derivation of w has length $2n - 1$.

Proof. Binary rules increment the length, and increment the variable count. Unary rules preserve the length, and decrement the variable count. Since w is comprised only of letters, exactly $n - 1$ binary rules and n unary rules were used. □

We will show that every context-free grammar is equivalent to a Chomsky normal form grammar, and there is an algorithm to produce such a grammar. There are three types of rules that are obstructions to a context-free grammar being in Chomsky normal form:

(i) rules $A \to \alpha$ where $|\alpha| \geq 2$ and α contains a letter;

(ii) rules of the form $A \to B$, called *unit rules*.

(iii) rules $A \to \alpha$ where $|\alpha| > 2$ and α contains only variables.

Suppose we have a rule of the form $A \to \alpha$ where $|\alpha| \geq 2$, and α contains a letter. For each letter $a \in \Sigma$, we can add a variable X_a and a rule $X_a \mapsto a$. Then we convert α to $X(\alpha)$, where X is the map converting each a into X_a. Then α contains no letter. We can therefore suppose without loss of generality that a given context-free grammar has no rules of this form.

Now consider a unit rule $A \to B$. A grammar is called *unit closed* if for all $A \to B \in P$ and $B \to \alpha \in P$, we also have $A \to \alpha \in P$. We can easily convert each grammar into an equivalent unit closed grammar by adding at most $|V| \cdot |P|$ new rules. If a context-free grammar G is unit closed, we will show that we can remove all unit rules to give a grammar G' without changing the language. Clearly $\mathcal{L}(G') \subseteq \mathcal{L}(G)$. Suppose $w \in \mathcal{L}(G)$, then w has a shortest G-derivation. Suppose this G-derivation of w contains a unit rule, so

$$S \overset{G}{\longrightarrow} \alpha A \beta \overset{G}{\longrightarrow}_1 \alpha B \beta \overset{G}{\longrightarrow} w$$

where this use of the unit rule is the last such usage. Since w contains no variables, we must have applied a rule $B \overset{G}{\longrightarrow}_1 \zeta$.

$$S \overset{G}{\longrightarrow} \alpha A \beta \overset{G}{\longrightarrow}_1 \alpha B \beta \overset{G}{\longrightarrow} \gamma B \delta \overset{G}{\longrightarrow}_1 \gamma \zeta \delta \overset{G}{\longrightarrow} w$$

where the $B \overset{G}{\longrightarrow}_1 \zeta$ is the first usage of a rule for B. Since $\alpha B \beta \overset{\gamma}{\to} B\delta$ did not use any B-rule by assumption, by unit closure we can replace this derivation with

$$S \overset{G}{\longrightarrow} \alpha A \beta \overset{G}{\longrightarrow} \gamma A \delta \overset{G}{\longrightarrow}_1 \gamma \zeta \delta \overset{G}{\longrightarrow} w$$

This is clearly shorter. So the shortest G-derivation contains no use of a unit rule, so is also a G'-derivation.

Finally, let us consider a rule $A \to \alpha$ where $|\alpha| > 2$ and α contains only variables. Suppose $\alpha = A_1 \dots A_n$. We define new variables $X_1, \dots, X_{n-2}$, and add new rules $A \to A_1 X_1, X_1 \to A_2 X_2, \dots X_{n-2} \to A_{n-1} A_n$. Then, performing this for all such rules, we obtain a grammar without any such rules. This grammar is in Chomsky normal form.

Theorem (Chomsky). Let G be a context-free grammar. Then we can compute an equivalent context-free grammar G' in Chomsky normal form.

Proof. Remove problems due to rules of the form $A \to \alpha$ where α contains a letter and has length at least 2. Form the unit closure, then remove unit rules. Remove problems due to rules of the form $A \to A_1 A_2 A_3 \dots A_n$. □

4.4. The pumping lemma for context-free languages

Definition. Let L be a context-free language, and let $n \in \mathbb{N}$. Suppose that for all $w \in L$ such that $|w| \geq n$, there are x, y, z, u, v such that $w = xuyvz$, and $|uyv| \leq n$, $|uv| > 0$, and for all k, $xu^k yv^k z \in L$. Then L satisfies the *pumping lemma for context-free languages with pumping number n*. We call u, v the *pump*.

Remark. The pump now has two parts, and one part may be the empty string. There is no longer a constraint on the position of the pump in a word; x and z may be of any length. The regular pumping lemma implies the context-free pumping lemma. Since there are uncountably many languages satisfying the regular pumping lemma, there are also uncountably many languages satisfying the context-free pumping lemma. In particular, the context-free pumping lemma cannot characterise the countable class of all context-free languages.

Theorem. Every context-free language satisfies the context-free pumping lemma for some pumping number n.

Proof. Let L be a context-free language. Then L has a Chomsky normal form grammar $G = (\Sigma, V, P, S)$, so $\mathcal{L}(G) = L$. Let $m = |V|$, and $n = 2^m$. We claim n is a pumping number for G.

If $\mathbb{T}$ is a G-parse tree where the height of $\mathbb{T}$ is $h + 1$ and $\sigma_{\mathbb{T}}$ is a word, then $|\sigma_{\mathbb{T}}| \leq 2^h$. Indeed, the largest possible tree of height $h + 1$ has 2^{h+1} leaves. Since $\sigma_{\mathbb{T}}$ is a word, we must have applied a unary rule for each letter. Every unary rule reduces the amount of leaves by one. Thus, the tree must contain $2^{h+1} - |\sigma_{\mathbb{T}}|$ leaves. Hence $|\sigma_{\mathbb{T}}| \leq 2^h$.

Consider a word $w \in \mathcal{L}(G)$ with $|w| \geq 2^m = n$. Then, if $\mathbb{T}$ is a G-parse tree of w, so $\sigma_{\mathbb{T}} = w$, we know by the previous claim that the height of $\mathbb{T}$ is at least $m + 1$. Let $t \in T$ such that the length of the branch to t is the height $h \geq m + 1$ of the tree. Then, the path from ε to t has $h + 1$ labels, so contains h variables and one letter.

Let s be an element of the branch of t such that the height of the subtree T_s is exactly $m + 1$. Hence $|\sigma_{\mathbb{T}_s}| \leq 2^m = n$. In particular, the path from s to t has $m + 2$ labels, so contains exactly

$m + 1$ variables and one letter. By the pigeonhole principle, there are two nodes $t_0 \subsetneq t_1$ on the branch from s to t with the same label $A \in V$. Let

$$\sigma_{\mathbb{T}} = a\sigma_{\mathbb{T}_s}b; \quad \sigma_{\mathbb{T}_s} = x'\sigma_{\mathbb{T}_{t_0}}z'; \quad \sigma_{\mathbb{T}_{t_0}} = u\sigma_{\mathbb{T}_{t_1}}v; \quad \sigma_{\mathbb{T}_{t_1}} = y$$

Then let $x = ax'$; $\quad z = z'b$ in the definition of the pumping lemma. Since $t_0 \neq t_1$, we have $|uv| > 0$. Note that $uyv = \sigma_{\mathbb{T}_{t_0}}$, which has length at most $|\sigma_{\mathbb{T}_s}|$, which has length at most $2^m = n$.

Pumping down is accomplished by grafting T_{t_1} into T_{t_0}; conversely, pumping up is accomplished by iteratively grafting T_{t_0} into T_{t_1}. Define $\mathbb{T}_{(0)} = \mathbb{T}_{t_1}$, and $\mathbb{T}_{(i+1)} = \mathrm{graft}(\mathbb{T}_{t_0}, t_1, \mathbb{T}_{(i)})$. Then $\mathbb{T}_k = \mathrm{graft}(\mathbb{T}, t_1, \mathbb{T}_{(k)})$, and $\sigma_{\mathbb{T}_k} = xu^kyv^kz$, thus proving the pumping lemma. □

Example. Consider the language $L = \{a^nb^nc^n \mid n > 0\}$ is not context-free. Suppose it is context-free. Then it has a pumping number $N \in \mathbb{N}$. Consider the word $a^Nb^Nc^N \in L$. Then $a^Nb^Nc^N = xuyvz$ where $|uyv| \leq N$, $|uv| > 0$. Since $|uyv| \leq N$, the string uyv cannot consist of all three letters a, b, c. In any case, pumping up the string will increase the quantity of one letter, but not increase the quantity of some other letter. Then the new word does not lie in L.

4.5. Closure properties

We have seen that $L = \{a^nb^nc^n \mid n > 0\}$ is not context-free. However, $L_0 = \left\{a^nb^nc^k \mid n, k > 0\right\}$ and $L_1 = \left\{a^kb^nc^n \mid n, k > 0\right\}$ are context-free, as the concatenation of context-free languages. But the intersection $L_0 \cap L_1$ is exactly L, so context-free languages are not closed under intersection. Therefore, they are also not closed under complement or difference, because this, along with closure under union, would imply closure under intersection. Note that any model of computation corresponding to context-free grammars cannot have a product construction, because such a construction would imply closure of context-free languages under intersection.

It can be shown that context-free languages correspond to *pushdown automata*, which are similar to deterministic automata, except that they also have a *stack*, which is a way of storing a sequence of arbitrary symbols. The stack has a *push* operation allowing a symbol to be pushed onto the top of the stack, and a *pop* operation that removes the topmost element currently on the stack. In particular, the stack does not have random access, and any symbol pushed can be popped at most once. Sequences that are pushed onto the stack are popped off in reverse order.

The transition function δ has an additional input denoting the topmost element currently on the stack, and an additional output describing an operation to perform on the stack, if any.

Theorem. A language is context-free if and only if there is a pushdown automaton for the language.

4.6. Decision problems

The word problem is already solved for context-free languages. The emptiness problem can be solved by the pumping lemma, similarly to the solution for regular languages. Indeed, if n is a pumping number, no word with length at most n can be the shortest word, since it can be pumped down. So we can explicitly check each word of length less than n to solve the emptiness problem. Note that the choice of pumping lemma to use does not matter in this argument.

5. Register machines

5.1. Definition

A register machine uses an alphabet Σ, has finitely many states, and finitely many *registers*, which are last-in first-out storage units containing a word $w \in \mathbb{W}$. The machine is able to access the last letter of the word, remove it, or push a new letter. A *configuration* or *snapshot* of length $n+1$ is a tuple of the form $(q, w_0, \dots, w_n) \in Q \times \mathbb{W}^{n+1}$. A configuration defines the state of the computation at a particular point in time.

The transition function should now be of the form $\delta : Q \times \mathbb{W}^{n+1} \to Q \times \mathbb{W}^{n+1}$. However, not every such function represents a real computation; there are uncountably many such functions, and the action on the registers is unrestricted.

Definition. Let Σ be an alphabet, and Q be a nonempty finite set of states. A tuple of the form

$$\begin{aligned}(0, k, a, q) &\in \mathbb{N} \times \mathbb{N} \times \Sigma \times Q \\ (1, k, a, q, q') &\in \mathbb{N} \times \mathbb{N} \times \Sigma \times Q \times Q \\ (2, k, q, q') &\in \mathbb{N} \times \mathbb{N} \times Q \times Q \\ (3, k, q, q') &\in \mathbb{N} \times \mathbb{N} \times Q \times Q\end{aligned}$$

is called a (Σ, Q)-*instruction*. For improved readability, we write

$$\begin{aligned}+(k, a, q) &= (0, k, a, q) \\ ?(k, a, q, q') &= (1, k, a, q, q') \\ ?(k, \varepsilon, q, q') &= (2, k, q, q') \\ -(k, q, q') &= (3, k, q, q')\end{aligned}$$

Intuitively,

- $+(k, a, q)$ represents pushing the letter a onto register k, then advancing to state q.
- $?(k, a, q, q')$ checks if the letter a is currently at the top of register k. If so, we advance to state q, and otherwise, we advance to state q'.
- $?(k, \varepsilon, q, q')$ checks if register k is empty. If so, we advance to state q, and otherwise, we advance to state q'.
- $-(k, q, q')$ pops the topmost letter from register k. If the register was already empty, we advance to state q, and otherwise, we advance to state q'.

These semantics are defined formally later. Let $\mathrm{Instr}(\Sigma, Q)$ be the set of (Σ, Q)-instructions. This is in principle an infinite set, but finite if k is bounded.

Definition. A tuple $M = (\Sigma, Q, P)$ is called a Σ*-register machine* if Σ is an alphabet, Q is a finite set of *states* with two distinguished states $q_S \neq q_H$, called the *start state* and *halt state* respectively, and $P : Q \to \operatorname{Instr}(\Sigma, Q)$ is the *program*. If $Q = \{q_0, \dots, q_n\}$, we can describe P as a finite collection of *program lines* $q_i \mapsto P(q_i)$. Since Q is finite, only finitely many registers k are referenced by P; we call the largest such k the *upper register index* of M.

Definition. Let M be a register machine with upper register index n and $\mathbf{w} = (w_0, \dots, w_n) \in \mathbb{W}^{n+1}$. For configurations C, C', we say M *transforms* C into C' if one of the following holds.

- $P(q) = +(k, a, q')$ and $C' = (q', w_0, \dots, w_{k-1}, w_k a, w_{k+1}, \dots, w_m)$.
- $P(q) = \ ?(k, a, q', q'')$, and
  - $w_k = wa$ for some w and $C' = (q', w_0, \dots, w_m)$, or
  - $w_k \neq wa$ for all w and $C' = (q'', w_0, \dots, w_m)$.
- $P(q) = \ ?(k, \varepsilon, q', q'')$, and
  - $w_k = \varepsilon$ and $C' = (q', w_0, \dots, w_m)$, or
  - $w_k \neq \varepsilon$ and $C' = (q'', w_0, \dots, w_m)$.
- $P(q) = -(k, q', q'')$, and
  - $w_k = \varepsilon$ and $C' = (q', w_0, \dots, w_m)$, or
  - $w_k = wa$ and $C' = (q'', w_0, \dots, w_{k-1}, w, w_{k+1}, \dots, w_m)$.

Then we define the *computation sequence* of M with input $\mathbf{w}$ by $C(0, M, \mathbf{w}) = (q_S, \mathbf{w})$, $C(k+1, M, \mathbf{w}) = C'$ where M transforms $C(k, M, \mathbf{w})$ into C'.

Remark. This recursive definition requires that the length of $\mathbf{w}$ is at least $n + 1$, where n is the upper register index. By convention, if $\mathbf{w}$ is too short, we pad it with copies of the empty word ε.

Remark. As defined above, all computation sequences are infinite, because every configuration is transformed by M into some other.

Definition. We say that the computation of M with input $\mathbf{w}$ *halts at time k* or *in k steps* if k is the smallest natural such that $C(k, M, \mathbf{w}) = (q_H, \mathbf{v})$. In this case, we say that $\mathbf{v}$ is the *register content at time of halting*, or the *output* of the computation. If such a k does not exist, we say the computation *does not halt.*

5.2. Strong equivalence

Definition. We say that register machines M, M' are *strongly equivalent* if for all k and $\mathbf{w}$, $C(k, M, \mathbf{w})$ and $C(k, M', \mathbf{w})$ have the same register content, and for all $\mathbf{w}$, we have that M halts after k steps with input $\mathbf{w}$ if and only if M' halts after k steps with input $\mathbf{w}$.

Remark. If $|Q| = |Q'|$, then for every (Σ, Q, P) there exists a strongly equivalent register machine (Σ, Q', P') by relabelling the states in P.

Proposition (the padding lemma)**.** Let M be a register machine. Then there are infinitely many different register machines that are strongly equivalent to M.

Proof. Let $M = (\Sigma, Q, P)$. The register machine completely determines the computation sequence, so after adding a new state $\hat{q}$ to Q, $\hat{q}$ is never a state in any computation sequence. So $(\Sigma, Q \cup \{\hat{q}\}, P \cup \{\hat{p}\})$ is strongly equivalent to M for any program line $\hat{p}$ for $\hat{q}$. □

Proposition. Up to strong equivalence, there are only countably many register machines.

Proof. Only the cardinality of Q matters up to strong equivalence. Let $M_{n,k}$ be the collection of register machines with a fixed state set with $|Q| = n$ and upper register index at most k. By checking cases, we find $|\mathrm{Instr}(\Sigma, Q)| = (k+1)n|\Sigma| + (k+1)n^2|\Sigma| + (k+1)n^2 + (k+1)n^2 = N_{n,k}$, which is finite. Therefore, there are $N_{n,k}^n$ different programs, and hence $|M_{n,k}| = N_{n,k}^n$ is finite. Then the collection of all register machines up to strong equivalence is $\bigcup_{n,k} M_{n,k}$ which is countable. □

5.3. Performing operations and answering questions

Definition. An *operation* is a partial function $f : \mathbb{W}^{n+1} \rightharpoonup \mathbb{W}^{n+1}$. We write $f(\mathbf{w}) \downarrow$ if $\mathbf{w}$ lies in the domain of f, and we say the operation *is defined* or *converges*. We write $f(\mathbf{w}) \uparrow$ otherwise, and say that the operation is *undefined* or *diverges*. A register machine M *performs* an operation f if for all $\mathbf{w}$, $f(\mathbf{w}) \downarrow$ if and only if M halts on input $\mathbf{w}$, and in this case, the register content at time of halting is $f(\mathbf{w})$.

Example. The operation 'never halt' is the empty function, $\operatorname{dom} f = \varnothing$. Then any program that never references the halt state in the right hand side of a program line performs this operation. For example, $q_S \mapsto +(0, a, q_S)$ and $q_H \mapsto +(0, a, q_S)$ suffices.

Remark. There are many register machines that perform the same operation, including many that are not strongly equivalent.

Example. The operation 'halt without doing anything' is the function $f(\mathbf{w}) = \mathbf{w}$ with $\operatorname{dom} f = \mathbb{W}^{n+1}$. An example of a program to perform this is $q_S \mapsto ?(0, a, q_H, q_H)$. This halts after one step, and preserves the register content.

Definition. A *question with* $k+1$ *answers* is a partition of $\mathbb{W}^{n+1}$ into $k+1$ sets A_i. A register machine *answers a question* if it has $k+1$ *answer states* $\overline{q}_i$, and upon input of $\mathbf{w}$, after finitely many steps its configuration is $(\overline{q}_i, \mathbf{w})$ for the value of i where $\mathbf{w} \in A_i$.

Example. The question 'is register i empty' is performed by $q_S \mapsto ?(i, \varepsilon, \overline{q}_Y, \overline{q}_N)$. The question 'is the final letter in register i the letter a' is performed by $q_S \mapsto ?(i, a, \overline{q}_Y, \overline{q}_N)$.

The following lemma allows us to concatenate register machines, or alternatively, to perform subroutines.

Lemma (concatenation). Let M perform $F : \mathbb{W}^{n+1} \rightharpoonup \mathbb{W}^{n+1}$, and M' perform $F' : \mathbb{W}^{n+1} \rightharpoonup \mathbb{W}^{n+1}$. Then we can construct a register machine $\hat{M}$ which performs $F' \circ F$.

Remark. If $F(\mathbf{w}) \uparrow$, then $(F' \circ F)(\mathbf{w}) \uparrow$. If $F(\mathbf{w}) \downarrow$ and $F'(F(\mathbf{w})) \uparrow$, then $(F' \circ F)(\mathbf{w}) \uparrow$. Otherwise, $(F' \circ F)(\mathbf{w}) \downarrow$.

Proof. We may assume without loss of generality that the state sets of the two machines are disjoint. We define $\hat{Q} = Q \cup Q' \setminus \{q_H\}$. We write $P^\star$ for the program P with the rule $q_H \mapsto P(q_H)$ removed, and then all instances of q_H replaced with q'_S. We then define $\hat{P} = P^\star \cup P'$. Then $\hat{M} = (\Sigma, \hat{Q}, \hat{P})$ clearly performs $F' \circ F$. □

Lemma (case distinction). Let Q be a question with $k + 1$ answers. Let $F_i : \mathbb{W}^{n+1} \rightharpoonup \mathbb{W}^{n+1}$ be operations for $i \leq k$. Let M be a register machine that answers Q, and let M_i be register machines that perform F_i. Then there is a register machine that performs the operation given by $G(\mathbf{w}) = F_i(\mathbf{w})$ if $\mathbf{w} \in A_i$.

Proof. We assume that Q is disjoint from each Q_i, and $\bigcap_{i \leq k} Q_i = \{q_H\}$. Let $P_i^\star$ be P_i where all occurrences of $q_{S,i}$ are replaced with the ith answer state $\overline{q}_i$. Define $Q^\star = Q \cup \bigcup_{i \leq k} Q_i \setminus \{q_{S,i}\}$ and $P^\star = P \cup \bigcup_{i \leq k} P_i^\star$. Then $M^\star = (\Sigma, Q^\star, P^\star)$ performs G. □

5.4. Register machine API

We can perform many different operations and answer many different questions using register machines. We say that a register is *unused* if no program line references it. A register is *empty* if it contains the empty word. Registers that are used only for computation and not the output are sometimes called *scratch space* or *scratch registers*.

- Consider
$$F(\mathbf{w}) = \begin{cases} \mathbf{w} & w_i \neq \varepsilon \\ \uparrow & w_i = \varepsilon \end{cases}$$
The question 'is register i empty' is performed by a register machine, and in this case, the 'never halt' operation can be performed; in the other case, the 'halt without doing anything' operation can be performed.
- The operation 'delete the final letter of register i if it exists' is performed by the program $q_S \mapsto -(i, q_H, q_H)$.
- The operation 'add letter a to register i' is performed by $q_S \mapsto +(i, a, q_H)$. Note that this machine also performs the operation 'guarantee that the ith register is nonempty'.
- The operation 'delete the content of register i' is performed by $q_S \mapsto -(i, q_H, q_S)$.
- We can perform the operation 'add a fixed word w to register i'. If $w = a_0 \dots a_\ell$, we use the concatenation lemma to perform the operation 'add letter a_j to register i' for each letter in the word.

- The operation 'replace the register content of i with the word w' can be performed by concatenating the operations 'delete the content of register i' and 'add w to register i'.
- We can answer the question 'what is the final letter of register i'. This question has $|\Sigma| + 1$ answers, since the register could be empty. For each letter $a_j \in \Sigma$, we ask the question 'does register i end in letter a_j', and if yes, go to the corresponding answer state $\overline{q}_j$, and if not, go to a state that asks the next question in the sequence. If no question answers 'yes', the register is empty, and we go to an answer state $\overline{q}_\varepsilon$.
- In particular, we can perform the operation 'copy the final letter of register i into register j if it exists', by asking what this letter is, and then in each case, pushing the relevant letter onto register j.
- We can also 'move the final letter of register i into register j if it exists' by first copying the letter and then removing the original from register i.
- The operation 'move the content of register i into register j in reverse order' is accomplished by repeatedly moving a single letter until no more letters lie in register i.
- The operation 'move the content of register i into register j in the correct order' can be performed by considering an unused empty register k. We move the register content from i to k in reverse order and then from k to j in reverse order.
- The operation 'reverse the content of register i' is performed by moving it in reverse order to an unused empty register j, and then moving this into i in the correct order.
- The operation 'move the content of register i into registers j and k in reverse order' is easily performed by copying the final letter of register i into j and then into k, then removing the final letter in register i iteratively until it is empty.
- The operation 'copy the content of register i into register j in reverse order' is accomplished by moving the content of register i into j and an unused empty register k, and then moving the register content of k into i in reverse order.
- The operation 'copy the content of register i into register j in the correct order' is accomplished by copying in the reverse order, and then reversing the content of register j.
- Consider the question 'is the content of register i the word w'. Let $w = a_0 \dots a_k$. We define the subroutine S_ℓ to answer the question 'is a_ℓ the final letter of register i'. If no, move to a state q_N. If yes, move the final letter to an unused empty register k and run subroutine $S_{\ell-1}$, or if $\ell = 0$, move to a state q_Y. At state q_N we move the content of k to i and answer $\overline{q}_N$, and at state q_Y we move the content of k to i and answer $\overline{q}_Y$.

6. Computability theory

6.1. Computable functions and sets

Remark. A lot of computations require the use of scratch space, and we want to reduce the mathematical information related to this scratch space. In the following definition, only register zero is considered real output; all other registers are considered scratch space.

Definition. Let M be a register machine, and let $k \in \mathbb{N}$. Then we define $f_{M,k} : \mathbb{W}^k \rightharpoonup \mathbb{W}$ by $f_{M,k}(\mathbf{w}) \uparrow$ when M does not halt on input $\mathbf{w}$, and $f_{M,k}(\mathbf{w}) = v_0$ when M halts on input $\mathbf{w}$ with halting register content $\mathbf{v}$.

Note that if M, M' are strongly equivalent, $f_{M,k} = f_{M',k}$ for all k. The converse does not hold. For the special case of $k = 1$, we also write $W_M = \operatorname{dom} f_{M,1}$.

Definition. A partial function $f : \mathbb{W}^k \rightharpoonup \mathbb{W}$ is called *computable* if there is a register machine M such that $f = f_{M,k}$.

Remark. There are only countably many computable functions, because there are only countably many register machines up to strong equivalence. For each computable function f, there are infinitely many register machines M such that $f = f_{M,k}$, since any register machine has infinitely many other strongly equivalent register machines. Due to the concatenation lemma and the case distinction lemma, computable functions are closed under concatenation and case distinction.

Example. The identity function on $\mathbb{W}$ is computable. Consider $c : \mathbb{W}^k \to \mathbb{W}$ is given by $c(\mathbf{w}) = v$ for a fixed v. The operation 'replace the content of register 0 with v' is performable on a register machine, so c is computable. The projection $\pi_i : \mathbb{W}^k \to \mathbb{W}$ given by $\pi_i(\mathbf{w}) = w_i$ is computable since the operation 'replace the content of register 0 with register i' can be performed on a register machine by emptying register 0 and then moving the content of register i to register 0.

Definition. Let $X \subseteq \mathbb{W}^k$. We say that a total function $f : \mathbb{W}^k \to \mathbb{W}$ is *a characteristic function of X* if $f(\mathbf{w}) \neq \varepsilon$ if and only if $\mathbf{w} \in X$. Let $a \in \Sigma$. We say that f is *the characteristic function of X* if $f(\mathbf{w}) = a$ if $\mathbf{w} \in X$ and $f(\mathbf{w}) = \varepsilon$ otherwise.

We use the notation χ_X for the characteristic function.

Definition. A set $X \subseteq \mathbb{W}^k$ is *computable* if the characteristic function χ_X of X is computable.

Note that a language is a set of words, so we can now reason about computability of languages.

Definition. Let $X \subseteq \mathbb{W}^k$. A partial function $f : \mathbb{W}^k \rightharpoonup \mathbb{W}$ is called *a pseudocharacteristic function of X* if $\operatorname{dom} f = X$. f is called *the pseudocharacteristic function of X* if $f(\mathbf{w}) = a$ if $\mathbf{w} \in X$, and undefined otherwise.

We use the notation ψ_X for the pseudocharacteristic function.

Definition. A set $X \subseteq \mathbb{W}^k$ is *computably enumerable* if the pseudocharacteristic function ψ_X is computable.

Remark. We will show that every computable set is computably enumerable, but the converse does not hold. We will also show that the computably enumerable sets are exactly the type 0 languages (those languages that have grammars), and that the class of computable languages is properly contained between type 1 and type 0.

6.2. Computability of languages

Proposition. Let $X \subseteq \mathbb{W}^k$. Then:

(i) X is computable if and only if X^c is computable.

(ii) X is computably enumerable if and only if there exists a register machine M such that $X = \operatorname{dom} f_{M,k}$.

(iii) If X is computable, then X is computably enumerable.

Proof. To simplify notation we consider the case $k = 1$. Note that if g and h are computable, then by the case distinction lemma, so is f defined by $f(w) = g(w)$ if $w \neq \varepsilon$, and $f(w) = h(w)$ if $w = \varepsilon$.

For the first part, consider the computable function f_1 given by $g(w) = \varepsilon$ and $h(w) = a$. Then $f_1 \circ \chi_X = \chi_{X^c}$, $f_1 \circ \chi_{X^c} = \chi_X$.

Now consider f_2 given by $g(w) = a$ and $h(w) = \varepsilon$. If $X = \operatorname{dom} f$, then $\psi_X = f_2 \circ f$.

Finally, consider f_3 given by $g(w) = a$ and $h(w) \uparrow$. Then $\psi_X = f_3 \circ \chi_X$. □

Theorem. Every regular language is computable.

Proof. Let L be such a regular language. Let $D = (\Sigma, Q, \delta, q_0, F)$ be a deterministic automaton such that $L = \mathcal{L}(D)$. The first step in our program is to reverse the content of register 0 into register 1, because register machines read words in the opposite order of deterministic automata. For each $q \in Q$, the register machine will have a set of states Q_q that indicate that we are currently mimicking D in state q. We will now move into the state set Q_{q_0}.

When moving into each state set Q_q, our program will read the final letter of register 1. If there are no letters in register 1, go to a fixed accepting state if $q \in F$ and the non-accepting state if $q \notin F$. Otherwise, let b be the last letter in register 1. Remove b from register 1, and go to state set $Q_{\delta(q,b)}$. We implicitly repeat this step, since we have now transitioned into a state set.

If the machine is in the given accepting state, we empty register 0, add a to register 0, and then halt. If the machine is in the non-accepting state, we empty register 0, and then halt. □

6.3. The shortlex ordering

We wish to create an order $<$ on $\mathbb{W}$ such that $(\mathbb{N}, <)$ is order-isomorphic to $(\mathbb{W}, <)$. We first fix an arbitrary total order $<$ on Σ.

Definition. The *shortlex ordering* on $\mathbb{W}$ given by an ordering of Σ is given by $w < v$ when

(i) $|w| < |v|$; or

(ii) $|w| = |v|$ but $w \neq v$, and for the least m such that the mth characters differ, the mth character of w is less than the mth character of v.

This ordering first checks length, then the lexicographic ordering. This is a total ordering on $\mathbb{W}$; it is irreflexive, transitive, and trichotomous. The empty word is the least element.

Example. Let $\Sigma = \{0, 1\}$, and fix $0 < 1$. Then an initial segment of the ordering is

$$\varepsilon, 0, 1, 00, 01, 10, 11, 000, 001, 010, 011, 100, 101, 110, 111, 0000, \dots$$

We can identify each word with a natural number, given by its index in this sequence, counting from zero. There are 2^k words of length k, so the index of the natural number associated to the word 0^k is exactly $2^k - 1$.

We can naturally extend the operations of addition and multiplication on the set of words by acting on the index of the word in this ordering. For example, $10 + 01 = 010$, because the associated index of 10 is 5, the index of 01 is 4, and the index of 010 is 9. This gives $\mathbb{W}$ the structure of a commutative semiring.

Theorem. The shortlex ordering has the same order type as $\mathbb{N}$. We write $(\mathbb{N}, <) \cong (\mathbb{W}, <)$.

Proof. For a fixed w, the set $\{v \mid v < w\}$ is finite. Therefore, the function $\# : \mathbb{W} \to \mathbb{N}$ given by $\#(w) = |\{v \mid v < w\}|$ is well-defined and is an order isomorphism. □

Theorem. The set $\{(v, w) \mid v < w\}$ is computable. The *successor function* $s : \mathbb{W} \to \mathbb{W}$ with $\#(s(w)) = \#(w) + 1$ is computable.

Proof. The question to determine the ordering of $|w_i|$ and $|w_j|$ can be answered by a register machine by copying i, j into empty registers and repeatedly removing letters until one or both is empty. If they have the same length, we again copy i, j into empty registers in the reverse order, and check and remove each letter until a difference is found.

To compute $s(w)$ for a word w, we find the last letter in w that is not the largest letter in the ordering. Replace this letter with the next letter in the ordering, and replace all subequent letters with the least letter in the ordering. If all k letters are the greatest letter, output the least letter $(k + 1)$-many times. □

6.4. Church's recursive functions

The class of recursive functions is defined inductively.

Definition. The *basic functions* are

$$\begin{aligned} \pi_{k,i} &\colon \mathbb{W}^k \to \mathbb{W} \\ c_{k,\varepsilon} &\colon \mathbb{W}^k \to \mathbb{W} \\ s &\colon \mathbb{W} \to \mathbb{W} \end{aligned}$$

where $\pi_{k,i}(\mathbf{w}) = w_i$, $c_{k,\varepsilon}(\mathbf{w}) = \varepsilon$, and $\#s(w) = \#w + 1$.

We call $\pi_{k,i}$ the *projection* functions, $c_{k,\varepsilon}$ the *constant* functions, and s the *successor* function.

Let $f \colon \mathbb{W}^m \rightharpoonup \mathbb{W}$ and $g_1, \dots, g_m \colon \mathbb{W}^k \rightharpoonup \mathbb{W}$. Then their *composition* is the function $h(\mathbf{w}) = f(g_1(\mathbf{w}), \dots, g_m(\mathbf{w}))$.

Let $f \colon \mathbb{W}^k \rightharpoonup \mathbb{W}$ and $g \colon \mathbb{W}^{k+2} \rightharpoonup \mathbb{W}$. Then the partial function $h \colon \mathbb{W}^{k+1} \rightharpoonup \mathbb{W}$ defined by $h(\mathbf{w}, \varepsilon) = f(\mathbf{w})$ and $h(\mathbf{w}, s(v)) = g(\mathbf{w}, v, h(\mathbf{w}, v))$ is a function defined by *recursion*.

Let $f \colon \mathbb{W}^{k+1} \rightharpoonup \mathbb{W}$. Then the function $h \colon \mathbb{W}^k \rightharpoonup \mathbb{W}$ defined by

$$h(\mathbf{w}) = \begin{cases} v & \text{if for all } u \leq v \text{, we have } f(\mathbf{w}, u) \downarrow \text{ and } v \text{ is } <\text{-minimal such that } f(\mathbf{w}, v) = \varepsilon \\ \uparrow & \text{if there is no } v \text{ satisfying the above property} \end{cases}$$

is a function defined by *minimisation*.

Remark. If a class of functions has the basic functions and is closed under composition, it has all constant functions $c_{k,v}(\mathbf{w}) = v$, because if $v = s^k(\varepsilon)$, $c_{k,v} = s^k \circ c_{k,\varepsilon}$.

Definition. A class $\mathcal{C}$ of partial functions is closed under composition, recursion, and minimisation if whenever $f_1, \dots, f_\ell \in \mathcal{C}$, then the results of applying these operations also lie in $\mathcal{C}$.

Remark. The class $\mathcal{P}$ of all partial functions is closed under composition, recursion, and minimisation.

Definition. We call a partial function *recursive* if it lies in the smallest class $\mathcal{C}$ that contains the basic functions and is closed under composition, recursion, and minimisation. A partial function is *primitive recursive* if it lies in the smallest class $\mathcal{C}$ that contains the basic functions and is closed under composition and recursion.

Example. $\pi_{1,0} \colon \mathbb{W}^1 \to \mathbb{W}$ is the identity function, which is primitive recursive. $\pi_{3,2} \colon \mathbb{W}^3 \to \mathbb{W}$ defined by $\pi_{3,2}(u, v, w) = w$ is primitive recursive as it is a basic function. The successor function $s \colon \mathbb{W} \to \mathbb{W}$ is primitive recursive. The function $s \circ \pi_{3,2}$ is primitive recursive, as the composition of primitive recursive functions.

The function h defined by $h(w, \varepsilon) = \pi_{1,0}(w)$ and $h(w, s(v)) = s \circ \pi_{3,2}(w, v, h(w, v)) = s(h(w, v))$ is primitive recursive, which is exactly the addition function $\#h(n, m) = \#n + \#m$.

We can define multiplication and exponentiation in a similar way, and so all of these are primitive recursive.

We can encode recursive functions in trees. Let T be a finitely branching tree, and define a labelling ℓ on T with the labels

	label	arity	branching number
projection	$B^{\pi}_{k,i}$	k	0
constant	$B^{c}_{k,i}$	k	0
successor	B^{s}	1	0
composition	$C_{n,k}$	k	$n+1$
recursion	R_k	$k+1$	2
minimisation	M_k	k	1

Definition. A tree T with a labelling ℓ is called a *recursion tree* if the branching of the tree corresponds exactly to the branching numbers of its labels, and

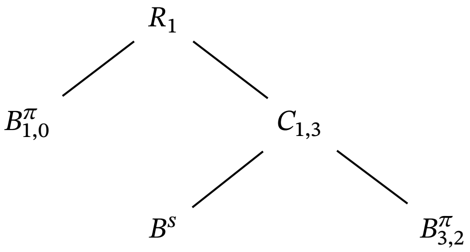
(i) if $\ell(s) = C_{n,k}$, then the first successor of s has a label of arity n and all other have labels with arity k;

(ii) if $\ell(s) = R_k$, then the first successor of s has arity k and the other has arity $k+2$;

(iii) if $\ell(s) = M_k$, then the successor has arity k.

A recursion tree is *primitive* if it has no minimisation labels M_k.

The following recursion tree describes the addition function defined above.



We can assign a (partial) recursive function $f_{T,\ell}$ to every recursive tree (T, ℓ). If the tree is primitive, the function obtained is primitive recursive.

Theorem. A partial function f is recursive if and only if there is a recursion tree (T, ℓ) such that $f = f_{T,\ell}$. It is primitive recursive if it admits a recursion tree that is primitive.

Proof. We can obtain the associated partial function from a recursion tree by induction on the height on the tree. For the converse, it suffices to show that the class of functions $f_{T,\ell}$ contains the basic functions and is closed under composition, recursion, and minimisation, which holds by construction. □

Theorem. Every partial recursive function is computable.

Proof. The basic functions have already been shown to be computable. Computable functions are closed under composition (previously called concatenation). So it suffices to show that the computable functions are closed under recursion and minimisation.

Let f, g be computable functions; we want to show that h defined by $h(\mathbf{w}, \varepsilon) = f(\mathbf{w})$ and $h(\mathbf{w}, s(v)) = g(\mathbf{w}, v, h(\mathbf{w}, v))$ is computable. We describe a register machine.

(i) Let k, ℓ be two empty unused registers.

(ii) Compute $f(\mathbf{w})$, and write the result to register ℓ. Note that if $f(\mathbf{w})$ is undefined, this produces the desired result.

(iii) If $v = \varepsilon$, output the content of register ℓ. Otherwise, apply the successor function s to register k and perform the following subroutine.

   (a) Compute $\mathrm{g}(\mathbf{w}, v, u)$ where u is the content of register ℓ, then overwrite register ℓ with the result.

   (b) Check whether v is equal to the register content of k. If so, output register ℓ. Otherwise, apply s to register k and restart the subroutine.

We now consider minimisation. Let f be computable. Let k be empty and unused. Perform the following subroutine.

(i) Compute $f(\mathbf{w}, u)$ where u is the content of register k. If this result is undefined, this is the desired result.

(ii) Check whether the computation result is empty. If it is empty, output the register content of k. Otherwise, apply the successor function s to k then restart the subroutine.

□

Remark. The proof showed that the computable functions are closed under recursion and minimisation, not just that all partial recursive functions are computable. Therefore, we can use recursion and minimisation directly to construct computable functions or register machines.

6.5. Merging and splitting words

There is a bijection $z : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$, called the *Cantor zigzag function*.

$$z(i, j) = \frac{(i + j)(i + j + 1)}{2} + j$$

This gives a bijection $\mathbb{W} \times \mathbb{W} \to \mathbb{W}$. All of these operations are computable by register machines.

Definition. Let v, w be words. Then we can *merge* the two words into $v * w$, which is the unique word such that $\#(v * w) = z(\#v, \#w)$. We can *split* a word w into u, v such that $\#w = z(\#u, \#v)$. We write $u = w_{(0)}$ and $v = w_{(1)}$.

Technically, splitting a word is not a computable function, since computable functions are defined to always have codomain $\mathbb{W}$. However, the operation of splitting a word can be performed.

6.6. Universality

Consider an alphabet Σ. We then have a notion of computability for sets $X \subseteq \Sigma^\star = \mathbb{W}$. If $\Sigma \subseteq \Sigma'$, then every Σ-register machine is a Σ'-register machine. However, the notion of Σ'-computability is no stronger than Σ-computability. One can show that computability over any alphabet Σ with $|\Sigma| \geq 2$ is equivalent to computability over the set $\{0, 1\}$ by encoding each letter as a binary string.

In this subsection, we aim to show that there is a *universal* register machine, which is a machine that can mimic every register machine. Let Σ be an alphabet, and add additional symbols

$$\mathbf{0}\,\mathbf{1} + - ?\,(\,)\,, \mapsto \square$$

We name the new alphabet Σ'. When we encode a mathematical object o as a word $\Sigma'^\star$, we write the encoded result $\mathrm{code}(o)$.

- We can encode $\mathbb{N}$ in binary using **0** and **1**, for instance, $\mathrm{code}(19) = \mathbf{10011}$.
- If $Q = \{q_0, \dots, q_k\}$, we define $\mathrm{code}(q_k) = \mathrm{code}(k)$.
- We encode instructions $I \in \mathrm{Instr}(\Sigma, Q)$ using $+ - ?\,,\,(\,)$; for instance, $\mathrm{code}(+(k, a, \ell)) = +(\mathrm{code}(k), a, \mathrm{code}(\ell))$.
- We encode program lines by $\mathrm{code}(q \mapsto I) = \mathrm{code}(q) \mapsto \mathrm{code}(I)$.
- We encode a register machine with program P as $\mathrm{code}(q_0 \mapsto P(q_0)), \dots, \mathrm{code}(q_n \mapsto P(q_n))$.
- We encode sequences of words by $\mathbf{w}$ by $\mathrm{code}(\mathbf{w}) = \square w_0 \square \dots \square w_k \square$.
- We encode configurations $(q, \mathbf{w})$ by $\mathrm{code}(q)\,\mathrm{code}(\mathbf{w})$.

Lemma. The function h defined by

$$h(w, u, v) = \begin{cases} \mathrm{code}(C(M, \mathbf{w}, \#v)) & \text{if } \exists M, \mathbf{w} \text{ such that } w = \mathrm{code}(M), u = \mathrm{code}(\mathbf{w}) \\ \uparrow & \text{otherwise} \end{cases}$$

is computable.

Proof. Define by recursion

$$h(\mathrm{code}(M), \mathrm{code}(\mathbf{w}), \varepsilon) = \mathrm{code}(q_0)\,\mathrm{code}(\mathbf{w}); \quad h(\mathrm{code}(M), \mathrm{code}(\mathbf{w}), s(v)) = \mathrm{code}(C')$$

where C' is the result of transforming $h(\mathrm{code}(M), \mathrm{code}(\mathbf{w}), v)$ by the machine M. □

Corollary. The *truncated computation* function $t_{M,k}$ defined by

$$t_{M,k}(\mathbf{w}, v) = \begin{cases} a & M \text{ has halted before time } \#v \text{ on input } \mathbf{w} \\ \varepsilon & \text{otherwise} \end{cases}$$

is computable.

Proof. Using recursion on the function h from the previous lemma, we check all values of h for words u such that $\#u < \#v$. If any of the values is in state q_H, output a, otherwise, output ε. □

Theorem (the software principle)**.** The function g defined by

$$g(v, u) = \begin{cases} f_{M,k}(\mathbf{w}) & \text{if } v = \text{code}(M), u = \text{code}(\mathbf{w}) \text{ and } \mathbf{w} \text{ has length } k \\ \uparrow & \text{otherwise} \end{cases}$$

is computable.

Proof. We have a computable function f that maps w, u, v to $\text{code}(C(M, \mathbf{w}, \#v))$ if $\text{code}(M) = w$ and $\text{code}(\mathbf{w}) = u$ by the previous lemma. We start by checking whether w is a code for a register machine and u is a code for a k-tuple of words; if not, never halt. Write f' for the computable function mapping w, u, v to a if the state of $f(w, u, v)$ is q_H, and ε otherwise. We minimise f' to obtain the computable function h, such that $h(w, u)$ is the least v such that $f(w, u, v)$ is in state q_H if it exists. If $h(w, u)$ does not halt, then there is no step at which the computation halts, as expected, since $g(w, u)$ should not halt in this case. If $h(w, u)$ halts, consider the configuration $C(M, \mathbf{w}, \#h(w, u))$ and find the code for its 0th register, and write this into the actual 0th register. □

Remark. A register machine U that computes g is called a *universal register machine*. U has a finite amount of used registers and states, but can mimic the behaviour of any register machine using an arbitrarily large amount of registers and states.

This allows us to streamline notation; for a word $v \in \mathbb{W}$, we can write

$$f_{v,k}(\mathbf{w}) = f_{U,2}(v, \text{code}(\mathbf{w})) = f_{M,k}(\mathbf{w})$$

if $\text{code}(M) = v$. Similarly, we can write $W_v = \text{dom} f_{v,1}$, so $\{W_v \mid v \in \mathbb{W}\}$ is the set of computably enumerable sets.

Theorem (*s*–*m*–*n* theorem; parameter theorem)**.** Let $g : \mathbb{W}^{k+1} \rightharpoonup \mathbb{W}$ be computable. Then there exists a total computable function $h : \mathbb{W} \to \mathbb{W}$ such that $f_{h(v),k}(\mathbf{w}) = g(\mathbf{w}, v)$.

This process is called *currying*, after Haskell Curry.

Remark. $g_v(\mathbf{w}) = g(\mathbf{w}, v)$ is a function in k variables. This is computable, so there is a mathematical function h such that $g_v = f_{h(x),k}$, but this h is not *a priori* computable.

Proof. First, the operation $\mathbf{w} \mapsto (\mathbf{w}, v)$ is performed by a register machine M_v; this is the register machine that writes v into register k. Therefore, we have a computable function $v \mapsto \mathrm{code}(M_v)$. Now, since g is computable, there is a register machine M such that $f_{M,k+1} = g$. Therefore, g_v is computed by the sequence of register machines M_v then M. We can computably concatenate two register machines, so we can compute a code for $M \circ M_v$. Hence the function $h(v) = \mathrm{code}(M \circ M_v)$ is total and computable.

We must show that $f_{h(v),k}(\mathbf{w}) = g(\mathbf{w}, v)$. Indeed,

$$f_{h(v),k}(\mathbf{w}) = f_{\mathrm{code}(M \circ M_v),k}(\mathbf{w}) = f_{M \circ M_v,k}(\mathbf{w}) = g_v(\mathbf{w}) = g(\mathbf{w}, v)$$

as required. □

6.7. The halting problem

Consider the sets

$$\mathbb{K}_0 = \{(w, v) \mid f_{w,1}(v)\downarrow\}; \quad \mathbb{K} = \{w \mid f_{w,1}(w)\downarrow\}$$

Theorem. $\mathbb{K}_0$ and $\mathbb{K}$ are computably enumerable.

Proof. It suffices to show that $\mathbb{K}_0, \mathbb{K}$ are the domains of computable functions. By the software principle, $f_{U,2}(w, v) = f_{w,1}(v)$ and $\mathrm{dom}\, f_{U,2} = \mathbb{K}_0$ as required. Observe that the diagonal function $\Delta(w) = (w, w)$ is computable, so $f_{U,2} \circ \Delta$ is computable, and $\mathrm{dom}(f_{U,2} \circ \Delta) = \mathbb{K}$. □

Theorem (the halting problem)**.** Neither $\mathbb{K}_0$ nor $\mathbb{K}$ are computable.

Proof. We prove the result for $\mathbb{K}_0$. Suppose that $\mathbb{K}_0$ is computable, so the characteristic function $\chi_{\mathbb{K}_0}$ is computable. Now, define

$$f(w) = \begin{cases} \uparrow & \text{if } \chi_{\mathbb{K}_0}(w, w) = a \\ \varepsilon & \text{if } \chi_{\mathbb{K}_0}(w, w) = \varepsilon \end{cases}$$

This is a computable function, so there is a machine $d \in \mathbb{W}$ such that $f_{d,1} = f$. Now,

$$f(d)\downarrow \iff f_{d,1}(d)\downarrow \iff (d, d) \in \mathbb{K}_0 \iff \chi_{\mathbb{K}_0}(d, d) = a \iff f(d)\uparrow$$

The proof is almost exactly the same for $\mathbb{K}$. □

6.8. Sets with quantifiers

Definition. $X \subseteq \mathbb{W}^k$ is called Σ_1 if there is a computable set $Y \subseteq \mathbb{W}^{k+1}$ such that $\mathbf{w} \in X \iff \exists y, (\mathbf{w}, y) \in Y$. We say $X = p(Y) = \{\mathbf{w} \mid \exists y,\ (\mathbf{w}, y) \in Y\}$ is the *projection* of Y. We say X is Π_1 if it is the complement of a Σ_1 set. We say X is Δ_1 if it is Σ_1 and it is Π_1.

Remark. The notation Σ is chosen to symbolise an existential quantifier, and Π symbolises the universal quantifier. In logic, sums and existentials are related, and products and universal quantifiers are also related. Δ is chosen for the German word *Durchschnitt* ('intersection'), as Δ_1 is the intersection of Σ_1 and Π_1.

Proposition. Every computable set is Δ_1.

Proof. By closure under complement, it suffices to show every computable set is Σ_1. The computable set $Y = \{(\mathbf{w}, y) \mid \mathbf{w} \in X\}$ has projection X. Logically, this adds a trivial existential quantification. □

Theorem. The computably enumerable sets are exactly the Σ_1 sets.

Proof. Suppose X is computably enumerable. Then by definition, the pseudocharacteristic function ψ_X is computable. Then there exists a register machine M such that $\psi_X = f_{M,k}$. We define $Y = \left\{(\mathbf{w}, y) \mid t_{M,k}(\mathbf{w}, y) = a\right\}$ where $t_{M,k}$ is the truncated computation function for the register machine M. Y is computable, since $t_{M,k} = \chi_Y$. Then $\mathbf{w} \in X \iff \psi_X(\mathbf{w}) \downarrow \iff \exists y,\ (\mathbf{w}, y) \in Y$ as required.

Now suppose X is Σ_1. Let Y be a computable set such that $X = p(Y)$. As the computable sets are closed under complement, the characteristic function χ_{Y^c} is computable. We apply minimisation to χ_{Y^c} to obtain a function h such that $h(\mathbf{w})$ is the minimal y such that $(\mathbf{w}, y) \in Y$. Then $\operatorname{dom} h = p(Y) = X$, so X is the domain of a partial computable function as required. □

Example. Let $f : \mathbb{W}^2 \rightharpoonup \mathbb{W}$ be a partial computable function in two variables. Then $X = \{w \mid \exists v,\ f(w, v) \downarrow\}$ is computably enumerable. Note that $f(w, v) \downarrow$ is not a computable predicate. Let M be a register machine such that $f = f_{M,2}$, and let

$$Z = \left\{(w, v_0, v_1) \mid t_{M,2}(w, v_0, v_1) = a\right\}$$

Clearly Z is computable. Define

$$Y = \left\{(w, u) \mid (w, u_{(0)}, u_{(1)}) \in Z\right\}$$

This is also computable. Now,

$$\begin{aligned} \exists v,\ f(w, v) \downarrow &\iff \exists v_0,\ \exists v_1,\ (w, v_0, v_1) \in Z \\ &\iff \exists u,\ (w, u_{(0)}, u_{(1)}) \in Z \\ &\iff (w, u) \in Y \\ &\iff w \in p(Y) \end{aligned}$$

So X is Σ_1 as required.

Remark. The previous argument is sometimes known as a *zigzag argument*; a pair of existential quantifiers can be merged into a single existential by merging the two words. Hence, we can perform infinitely many computations in parallel.

Corollary. The computable sets are exactly the Δ_1 sets.

Proof. If X is computable, it must be Δ_1 by a previous result. If X is Δ_1, we can use a zigzag technique. We know that there are machines M, M' such that $w \in X \iff \exists v,\ t_{M,k}(\mathbf{w}, v) = a$ and $w \notin X \iff \exists v,\ t_{M',k}(\mathbf{w}, v) = a$. Now, consider

$$f(\mathbf{w}, v) = \begin{cases} t_{M,k}(\mathbf{w}, v_{(1)}) & \# v_{(0)} \text{ is even} \\ t_{M',k}(\mathbf{w}, v_{(1)}) & \# v_{(0)} \text{ is odd} \end{cases}$$

This is computable. Apply minimisation to f to obtain a function h where $h(\mathbf{w})$ is the least v such that $f(\mathbf{w}, v) \neq \varepsilon$. We output a if $\# h(\mathbf{w})_{(0)}$ is even, and ε if $\# h(\mathbf{w})_{(0)}$ is odd. □

Corollary. Σ_1 is not closed under complement.

Proof. The complement of the halting set $\mathbb{W} \setminus \mathbb{K}$ is Π_1 and not Δ_1, so not Σ_1. □

Theorem. Every type 0 language is computably enumerable.

Proof. Let $G = (\Sigma, V, P, S)$ and let $\Sigma' = \Omega \cup \{\to\}$. We encode derivations as $\sigma_0 \to \cdots \to \sigma_n$; this is a Σ'-word. We say $w \in (\Sigma')^\star$ is a *derivation code* if w is of this form with $(\sigma_0, \dots, \sigma_n)$ a G-derivation. In this case, we call σ_0 the *initial string* and σ_n the *final string*. Let

$$Y = \{(w, v) \mid v \text{ is a derivation code with initial string } S \text{ and final string } w\}$$

Y is computable since we can produce a register machine that tests if a given derivation code can be produced from a fixed given grammar. But $w \in \mathcal{L}(G) \iff \exists v,\ (w, v) \in Y$. This is Σ_1, as required. □

Remark. The converse also holds; every computably enumerable set $X \subseteq \mathbb{W}$ is a type 0 language. This will not be proven rigorously in this course; a sketch will be provided later.

6.9. Closure properties

Proposition. The computable sets are closed under intersection, union, complement, difference, and concatenation.

Proof. Let A, B be computable sets, so χ_A, χ_B are computable functions. We obtain

$$\chi_{A \cap B}(\mathbf{w}) = \begin{cases} a & \chi_A(\mathbf{w}) = a \text{ and } \chi_B(\mathbf{w}) = a \\ \varepsilon & \text{otherwise} \end{cases}$$

For complement,

$$\chi_{\mathbb{W} \setminus A}(\mathbf{w}) = \begin{cases} a & \chi_A(\mathbf{w}) = \varepsilon \\ \varepsilon & \text{otherwise} \end{cases}$$

For concatenation, we suppose $A, B \subseteq \mathbb{W}$ are one-dimensional. Given a word w, we can iterate over all possible decompositions $w = vu$ and check if $v \in A, u \in B$. There are $(|w| + 1)$-many such decompositions, so this minimisation will always halt. □

Remark. The result for intersection is analogous to the product construction from deterministic automata; two computable functions can be evaluated in parallel since they always terminate, and then their results may be combined.

Proposition. The computably enumerable sets are closed under intersection, union, and concatenation. They are not closed under complement or difference.

Proof. We have already shown that the complement of the halting set $\mathbb{K}$ is Π_1 but not Σ_1, so the computably enumerable sets are not closed under complement or difference. For intersection, the same construction as before works.

$$\chi_{A \cap B}(\mathbf{w}) = \begin{cases} a & \psi_A(\mathbf{w}) = a \text{ and } \psi_B(\mathbf{w}) = a \\ \uparrow & \text{otherwise} \end{cases}$$

This is because if ψ_A or ψ_B diverge, the result is ↑ as desired. For union, we cannot compute ψ_A and ψ_B serially, since if $\psi_A \uparrow$ we never run ψ_B at all. Using the zigzag technique, we can check $\psi_A(\mathbf{w})$ and $\psi_B(\mathbf{w})$ in parallel, halting if either halts at any time index. This idea is elaborated on an example sheet.

For concatenation, consider the set Z of triples (w, v, u) such that v is an initial segment of w, and after $\#u$ steps, $\psi_A(v) = a$ and $\psi_B(v') = a$, where $w = vv'$. Now define $Y = \{(w, u) \mid (w, u_{(0)}, u_{(1)}) \in Z\}$, so $w \in AB$ if and only if there exists v such that $(w, v) \in Y$. □

Proposition. X is computably enumerable if and only if there is a partial computable function f such that $X = \operatorname{Im} f$.

Remark. In fact, a stronger result is true: X is computably enumerable if and only if there is a *total* computable function f such that $X = \operatorname{Im} f$. This is seen on an example sheet. This result justifies the name 'computably enumerable'.

Proof. If ψ_X is computable, then so is

$$f(w) = \begin{cases} w & \psi_X(w) \downarrow \\ \uparrow & \text{otherwise} \end{cases}$$

Clearly $\operatorname{Im} f = X$ as required.

Conversely, suppose $f : \mathbb{W} \rightharpoonup \mathbb{W}$ with $X = \operatorname{Im} f$. Suppose $f = f_{c,1}$. We use the zigzag technique. Define the set Z of tuples (w, v, u) such that $t_{c,1}(v, u) = a$ and $f_{c,1}(v) = w$. Let $Y = \{(w, v) \mid (w, v_{(0)}, v_{(1)}) \in Z\}$, so $\operatorname{Im} f = p(Y)$. □

6.10. The Church–Turing thesis

Register machines and recursive functions can both be used to define computability. Historically, *Turing machines* were also used to define and analyse computability. There is another alternative, known as *while programs*. Notably, in this model, there is no special 'halt state'; the program halts simply when there are no more instructions to execute. Therefore the computation sequence in this model may be finite. This gives rise to a notion of while computable functions, the functions computed by a while program.

Theorem. Let $f : \mathbb{W}^k \rightharpoonup \mathbb{W}$. Then, the following are equivalent.

(i) f is (register machine) computable.

(ii) f is partial recursive.

(iii) f is Turing computable.

(iv) f is while computable.

Turing machines, register machines, recursive functions, and while programs are superficially completely different approaches, yet the classes of computable functions that they define are exactly identical. The *Church–Turing thesis* is that this is universal; any reasonable notion of computation is equivalent. Unfortunately, this is a nonmathematical statement, and cannot be made precise; this is simply a statement that describes our intuition about what computation means. Accepting this thesis allows us to freely choose which notion of computability we would like to use for a given task.

The following is a proof sketch of the fact that computably enumerable sets are type 0 languages. The sketch makes use of the fact that Turing computability is exactly register machine computability. For more detail, see *Formal Languages* (Salomaa 1973).

Proof sketch. Let M be a Turing machine computing ψ_X. Without loss of generality, let the read-write head be then moved to the front, so $q_s\square w\square \xrightarrow{M} q_H\square a\square$. This is a rewrite system with the rules described by the definition of the Turing machine, transforming $q_S\square w\square$ into $q_H\square a\square$

We define a grammar which starts from S, with $S \to q_H\square a\square$, and performs all Turing instructions backwards. When q_S is seen, it deletes everything except w. □

6.11. Solvability of decision problems

We can use the Church–Turing thesis to give precise statements of our decision problems, without relying on an informal notion of 'algorithm'. First, we encode grammars in such a way that for all $w \in \mathbb{W}$, there exists a grammar G such that $\mathrm{code}(G) = w$; we write G_w for the associated grammar for a word. We require that all grammars are of the form G_w for some word $w \in \mathbb{W}$. Now,

(i) the word problem is $\{(w, v) \mid w \in \mathcal{L}(G_v)\}$;

(ii) the emptiness problem is $\{w \mid \mathcal{L}(G_w) = \varnothing\}$;

(iii) the equivalence problem is $\{(w, v) \mid \mathcal{L}(G_w) = \mathcal{L}(G_v)\}$.

These are sets of tuples of words, so we can use our notion of computability. We can now concretely define that such a problem is *solvable* if the set is computable.

Theorem. The word problem for type 0 grammars is unsolvable.

Proof. Let $W = \{(w, v) \mid w \in \mathcal{L}(G_v)\}$. We want to show that W is not computable. Recall that $\mathbb{K}_0 = \left\{(w, v) \mid f_{w,1}(v) \downarrow\right\}$; we will use a proof analogous to the one used for this set. Suppose W is computable, so let

$$f(w) = \begin{cases} \uparrow & w \in \mathcal{L}(G_w) \\ a & w \notin \mathcal{L}(G_w) \end{cases}$$

Then f is a computable function. Hence, $\operatorname{dom} f$ is computably enumerable. So there exists a grammar G such that $\mathcal{L}(G) = \operatorname{dom} f$. Let $d \in \mathbb{W}$ be such that $G = G_d$. Then

$$d \in \mathcal{L}(G_d) \iff d \in \operatorname{dom} f \iff d \notin \mathcal{L}(G_d)$$

□

6.12. Reduction functions

Definition. Let $A, B \subseteq \mathbb{W}$. A function $f \colon \mathbb{W} \to \mathbb{W}$ is called a *reduction* from A to B if f is total computable and $w \in A$ if and only if $f(w) \in B$. We write $A \leq_m B$ if there is a reduction from A to B.

Remark. Given a reduction f from A to B, the set A is intuitively 'at most as complicated as B'. Note that $f^{-1}(B) = A$.

The subscript m in the notation $A \leq_m B$ stands for 'many-one'; the function f need not be injective. Note that $\leq_m$ is reflexive and transitive. This relation respects complements: $A \leq_m B$ implies $\mathbb{W} \setminus A \leq_m \mathbb{W} \setminus B$. The relation is not in general antisymmetric, so this does not form a partial order. Instead, $\leq_m$ forms a (partial) preorder.

If $\leq$ is a preorder on a set X, we can define the equivalence relation $x \sim y$ when $x \leq y$ and $y \leq x$. Then $\left({}^{X}\!/\!_{\sim}, \leq\right)$ is a partial order. A preorder can therefore be understood as a partial order, except that instead of ordering single elements, it orders clusters of equivalent elements.

If $A \leq_m B$ and B is computable, then A is also computable. Similarly, if $A \leq_m B$ and B is computably enumerable, then A is also computably enumerable. This demonstrates the fact that $\chi_A = \chi_B \circ f$ and $\psi_A = \psi_B \circ f$, where f is the reduction.

Note that if $A \leq_m B$ and A is not computable, then B is also not computable, and a similar result holds for sets that are not computably enumerable. In particular, if $\mathbb{K} \leq_m A$, then A is not computable. If $\mathbb{W} \setminus \mathbb{K} \leq_m A$, then A is not computably enumerable.

Remark. Many of the previous proofs in this section have implicitly used the notion of a reduction function, for instance, the claim that solvability of the set $\{(w, v) \mid w \in \mathcal{L}(G_v)\}$ is equivalent to solvability of the set $\{(w, v) \mid w \in W_v\}$.

Proposition. Let A be a computable set, and $B \neq \varnothing, \mathbb{W}$. Then $A \leq_m B$.

Proof. Since $B \neq \varnothing, \mathbb{W}$, let $v \in B, u \notin B$. Since A is computable, we have the computable function

$$f(w) = \begin{cases} v & w \in A \\ u & w \notin A \end{cases}$$

This is a reduction from A to B as required. □

Note that $\mathbb{W}\backslash\mathbb{K} \nleq_m \mathbb{K}$, otherwise $\mathbb{K}$ is not computably enumerable. We also have $\mathbb{K} \nleq_m \mathbb{W}\backslash\mathbb{K}$ from the first result, after considering complements. There are therefore various different *degrees of unsolvability*: equivalence classes of $\leq_m$ that are strictly larger than the class of computable sets.

There are many more such classes than the ones containing $\mathbb{K}$ and $\mathbb{W} \setminus \mathbb{K}$. Let $\{0, 1\} \subseteq \Sigma$. If A, B are sets, we can define the *Turing join* $A \oplus B = 0A \cup 1B$. Then $A \leq_m A \oplus B$ and $B \leq_m A \oplus B$. The Turing join produces an upper bound in the set of equivalence classes of sets of words, and it can be shown that this is the least upper bound. Hence we obtain another class of sets represented by $\mathbb{K} \oplus \mathbb{W} \setminus \mathbb{K}$. This is neither Σ_1 nor Π_1.

Definition. If $\mathcal{C}$ is a class of sets, we say that A is *$\mathcal{C}$-hard* if for all $B \in \mathcal{C}$, we have $B \leq_m A$. We say that A is *$\mathcal{C}$-complete* if it is $\mathcal{C}$-hard and $A \in \mathcal{C}$.

Remark. A $\mathcal{C}$-hard set is 'at least as hard as $\mathcal{C}$'. A $\mathcal{C}$-complete set is the 'most complicated' $\mathcal{C}$ set.

Corollary. Let A be Δ_1 and $A \neq \varnothing, \mathbb{W}$. Then A is Δ_1-complete.

Proof. The Δ_1 sets are the computable sets, so we simply apply the previous proposition. □

Theorem. $\mathbb{K}$ is Σ_1-complete.

Proof. Clearly $\mathbb{K} \in \Sigma_1$. Now, let X be an arbitrary set in Σ_1, so X is computably enumerable. Let f be a partial computable function such that $X = \operatorname{dom} f$. It suffices to show $X \leq_m \mathbb{K}$.

Consider the function $g(w, u) = f(w)$. This is computable. We can therefore apply the *s–m–n* theorem to obtain a total computable function h such that $f_{h(w)}(u) = g(w, u) = f(w)$. We claim that h is a reduction function from X to $\mathbb{K}$.

Suppose $w \in X$. Then $w \in \operatorname{dom} f$, so $f_{h(w)}$ is the constant function $f(w)$. Hence $W_{h(w)} = \mathbb{W}$. So f is total, and therefore $f_{h(w)}(h(w)) \downarrow$. So $h(w) \in \mathbb{K}$.

Now suppose $w \notin X$, so $w \notin \operatorname{dom} f$. Then $f_{h(w)}$ does not halt for any input, giving $W_{h(w)} = \varnothing$. So $f_{h(w)}(h(w)) \uparrow$ and in particular $h(w) \notin \mathbb{K}$. □

6.13. Rice's theorem

We say that M and M' are weakly equivalent when $\operatorname{dom} f_{M,1} = W_M = W_{M'} = \operatorname{dom} f_{M',1}$. We can extend this to words. Words v, u are weakly equivalent when $W_v = W_u$, and write $v \sim u$.

Definition. A set $I \subseteq \mathbb{W}$ is called an *index set* if it is closed under weak equivalence.

Remark. Index sets are unions of equivalence classes.

Example. $\varnothing$ and $\mathbb{W}$ are the trivial index sets. Other index sets correspond to properties of computably enumerable sets. $\mathbf{Emp} = \{v \mid W_v = \varnothing\}$, $\mathbf{Fin} = \{v \mid W_v \text{ finite}\}$, $\mathbf{Inf} = \{v \mid W_v \text{ infinite}\}$, $\mathbf{Tot} = \{v \mid W_v = \mathbb{W}\}$ are index sets. Note that the emptiness problem is precisely the index set **Emp**.

Theorem (Rice's theorem)**.** No nontrivial index set is computable.

Fix $w \in \mathbb{W}$ and consider the function

$$g(u, v) = \begin{cases} f_{w,1}(v) & f_u(u) \downarrow \text{ or equivalently, } u \in \mathbb{K} \\ \uparrow & \text{otherwise} \end{cases}$$

This is computable, even though the case distinction itself is not computable. By the *s–m–n* theorem, there is a total function h such that

$$f_{h(u)}(v) = g(u, v) = \begin{cases} f_{w,1}(v) & u \in \mathbb{K} \\ \uparrow & u \notin \mathbb{K} \end{cases}$$

If $u \in \mathbb{K}$, then $W_{h(u)} = W_w$. If $u \notin \mathbb{K}$, then $W_{h(u)} = \varnothing$. This h will be used as a reduction function.

Proof. Let I be an index set. Let e be such that $W_e = \varnothing$. Then either $e \in I$, or $e \notin I$.

Suppose $e \in I$. Since I is nontrivial, there exists $w \notin I$, so $W_w \neq \varnothing$. Consider the function g from the discussion above, instantiated with this choice of w, and apply the *s–m–n* theorem to obtain a total function h. We claim that h reduces $\mathbb{W} \setminus \mathbb{K}$ to I. If $u \in \mathbb{K}$, then $W_{h(u)} = W_w$. Hence $h(u) \sim w$, so $h(u) \notin I$. If $u \notin \mathbb{K}$, then $W_{h(u)} = \varnothing$, so $h(u) \sim e$, so $h(u) \in I$.

Now suppose $e \notin I$. Then there exists $w \in I$, and $W_w \neq \varnothing$. Take g and h as before. We claim now that h reduces $\mathbb{K}$ to I. If $u \in \mathbb{K}$, then $W_{h(u)} = W_w$, so $h(u) \sim w$, so $h(u) \in I$. If $u \notin \mathbb{K}$, then $W_{h(u)} = \varnothing$, so $h(u) \sim e$, giving $h(u) \notin I$. □

Remark. The proof given for Rice's theorem shows a stronger statement: if $e \in I$ then $\mathbb{W} \setminus \mathbb{K} \leq_m I$, and if $e \notin I$ then $\mathbb{K} \leq_m I$. This allows us to show that certain index sets are not computably enumerable. $e \in$ **Emp** so $\mathbb{W} \setminus \mathbb{K} \leq_m$ **Emp**. Similarly, $\mathbb{W} \setminus \mathbb{K} \leq_m$ **Fin**. For the other two index sets, we can only deduce that $\mathbb{K} \leq_m$ **Inf** and $\mathbb{K} \leq_m$ **Tot**, since e does not lie in these sets.

Corollary. **Emp**, **Fin**, **Inf**, **Tot** are not computable.

Corollary. The emptiness problem for type 0 grammars is unsolvable.

Corollary. The equivalence problem for type 0 grammars is unsolvable.

Proof. We define $\mathbf{Eq} = \{(w, v) \mid W_w = W_v\}$. The function $g(w) = (w, e)$ can be performed by a register machine for any e. If $W_e = \varnothing$, then $\chi_{\mathbf{Emp}} = \chi_{\mathbf{Eq}} \circ g$. Hence, if **Eq** is computable, so is **Emp**. □

Remark. One can show that **Emp** is many-one equivalent to $\mathbb{W} \setminus \mathbb{K}$, so it is Π_1-complete, as proven on the last example sheet. The other problems **Tot**, **Inf**, **Fin** are not in Σ_1 or Π_1.

Theorem. **Fin** is not Σ_1 or Π_1.

Proof. We know $\mathbb{W} \setminus \mathbb{K} \leq_m \mathbf{Fin}$ by the proof of Rice's theorem, so **Fin** is not Σ_1. To show it is not Π_1, one must show that $\mathbb{K} \leq_m \mathbf{Fin}$. Consider

$$g(w, v) = \begin{cases} \uparrow & t_{w,1}(w, v) = a \\ \varepsilon & \text{otherwise} \end{cases}$$

Applying the *s–m–n* theorem, we obtain a total function h such that $f_{h(w),1}(v) = g(w, v)$. We show h reduces $\mathbb{K}$ to **Fin**. If $w \in \mathbb{K}$, then $f_{h(w),1}$ is undefined from v onwards, where v is the halting time of $f_w(w)$. Hence, $W_{h(w)}$ is finite, so $h(w) \in \mathbf{Fin}$. If $w \notin \mathbb{K}$, then $g(w, v) = \varepsilon$ for all v. So $f_{h(w),1}$ is the constant function with value ε. Hence $W_{h(w)} = \mathbb{W}$ which is infinite, so $h(w) \notin \mathbf{Fin}$. □

V. Galois Theory

Lectured in Michaelmas 2022 by PROF. A. J. SCHOLL

Suppose K and L are fields, and $K \subseteq L$. We can view L as a vector space over K, and therefore analyse things like its dimension. We study how these extensions of fields interact, and how they can embed inside each other.

To analyse these field extensions, we will understand the Galois group associated to a particular field extension $K \subseteq L$. This group describes the different ways that L can embed into itself, while preserving the structure of K. This turns out to provide a measure of the complexity of L with respect to K.

The central result of the course is the Galois correspondence. If $K \subseteq L$, there may be other fields lying between K and L. We prove that the subgroups of the Galois group correspond exactly to these intermediate fields.

We apply Galois theory to some problems that had been unsolved for many centuries or millenia. Classic examples include doubling the cube and trisecting the angle. We also prove that there is no formula for finding a root of the general quintic.

Contents

1. Polynomials

1.1. Introduction

Galois theory concerns itself with solving polynomial equations of higher degree, and discussing how the symmetries of these polynomials relate to their solubility. The modern interpretation of Galois theory is more interested in the fields that particular polynomials generate, rather than their particular solutions; this naturally extends to studying symmetries of fields.

1.2. Solving quadratics, cubics and quartics

Methods for solving quadratic equations have been known since the time of the Babylonians. Consider $aX^2 + bX + c$, and complete the square into $\left(X + \frac{1}{2}b\right)^2 + c - \frac{b^2}{4}$. This leads directly into the usual formula.

Alternatively, consider $(X - x_1)(X - x_2)$ and expand, giving $X^2 - (x_1 + x_2)X + x_1x_2$. Thus, $x_1 + x_2 = -b$ and $x_1x_2 = c$. We can write $x_1 = \frac{1}{2}[(x_1 + x_2) + (x_1 - x_2)]$, where $x_1 + x_2 = b$ and $(x_1 - x_2)^2 = b^2 - 4c$.

Cubics were solved much later, in the early 16th century, by the Italian mathematician del Ferro. Consider the cubic $X^3 + aX^2 + bX + c$, written as $(X - x_1)(X - x_2)(X - x_3)$. Multiplying, we find

$$x_1 + x_2 + x_3 = -a; \quad x_1x_2 + x_2x_3 + x_3x_1 = b; \quad x_1x_2x_3 = -c$$

Without loss of generality we can set $a = 0$ by replacing $X \mapsto X - \frac{a}{3}$. Now,

$$x_1 = \frac{1}{3}\left[(x_1 + x_2 + x_3) + \underbrace{(x_1 + \omega x_2 + \omega^2 x_3)}_{u} + \underbrace{(x_1 + \omega^2 x_2 + \omega x_3)}_{v}\right]$$

where $\omega = e^{\frac{2\pi i}{3}}$. The u, v are known as Lagrange resolvents. Applying a cyclic permutation to x_1, x_2, x_3 in u or v, we find $u \mapsto \omega u$ and $v \mapsto \omega v$. Hence, the cubes of u and v are invariant under cyclic permutations of x_1, x_2, x_3. Under a permutation $x_2 \mapsto x_3, x_3 \mapsto x_2$, u and v swap. Hence, $u^3 + v^3$ and u^3v^3 are invariant under all permutations of roots. A general fact that we will prove later is that such invariant expressions can be written in terms of the coefficients of the polynomial. In this case, we have

$$u^3 + v^3 = -27c; \quad u^3v^3 = -27b^2$$

Now, u^3 and v^3 are the roots of the quadratic $Y^2 + 27cY - 27b^2$. This then provides a formula for the root x_1. This process is known as Cardano's formula.

Similarly, the quartic $X^4 + aX^3 + bX^2 + cX + d$ can be solved by producing an auxiliary cubic equation, in a similar way to the auxiliary quadratic equation found for the cubic case above. However, the same process does not work for the quintic; the auxiliary equation has a degree which is too large. The underlying reason behind this is to do with group theory, and in particular, the group structure of S_5 and A_5. This will be explored later in the course.

1.3. Polynomial rings

In this course, *ring* means a commutative nonzero ring. If R is a ring, $R[X]$ denotes the ring of polynomials with elements $\sum_{i=0}^n a_i X^i$, and the usual operations of addition and multiplication. A polynomial $f \in R[X]$ can be interpreted as a function $f : R \to R$, given by $x \mapsto \sum_{i=0}^n a_i x^i$. It is, however, important to distinguish the polynomial and its associated function; the polynomial is not in general uniquely determined by the function. For example, let $R = {}^{\mathbb{Z}}\!/_{p\mathbb{Z}}$, so for all $a \in R$, we have $a^p = a$, and hence X^p and X are different polynomials yet represent the same function.

Recall from Groups, Rings and Modules that if $R = K$ is a field, $K[X]$ is a Euclidean domain (and hence is a unique factorisation domain, a Noetherian ring, a principal ideal domain, and an integral domain). Hence, there is a division algorithm: for polynomials $f, g \in K[X]$, there exists a unique $q, r \in K[X]$ such that $f = gq + r$ and $\deg r < \deg g$, where we denote $\deg 0 = -\infty$. If $g = X - a$ is linear, $f = (X - a)q + r$ where $r = f(a) \in K$; this is the familiar remainder theorem. Note that every polynomial $f \in K[X]$ is a product of irreducible polynomials since $K[X]$ is a unique factorisation domain, and there are greatest common divisors which can be computed using Euclid's algorithm in the usual way.

Proposition. Let K be a field, and $0 \neq f \in K[X]$. Then, f has at most $\deg f$ roots in K.

Proof. If f has no roots, the proof is complete. If f has a root a, consider $f = (X-a)q + f(0) = (X - a)q$. For a root b of f, either $b = a$ or $q(b) = 0$. By induction, q has at most $\deg q$ roots, since $\deg q < \deg f$. Then $\deg q + 1 \leq \deg f$ as required. □

1.4. Symmetric polynomials

Definition. Let R be a ring, and let $n \geq 1$. A polynomial $f \in R[X_1, \dots, X_n]$ is *symmetric* if, for every permutation $\sigma \in S_n$, we have $f(X_{\sigma(1)}, \dots, X_{\sigma(n)}) = f(X_1, \dots, X_n)$, where S_n is the symmetric group of degree n.

Note that constant polynomials are symmetric, and the property of symmetry is closed under addition and multiplication. Hence, the set of symmetric polynomials is a subring of $R[X_1, \dots, X_n]$.

Example. $X_1 + \dots + X_n$ is symmetric. More generally, $p_k = X_1^k + \dots + X_n^k$ is symmetric.

Proposition. Let $f\sigma(X) = f(X_{\sigma(1)}, \dots, X_{\sigma(n)})$. This gives an action (on the right) of the group S_n on $R[X_1, \dots, X_n]$. A polynomial $f \in R[X_1, \dots, X_n]$ is symmetric if and only if f is fixed under the action of S_n; in other words, $f\sigma = f$ for all $\sigma \in S_n$.

Definition. The *elementary symmetric functions* or *elementary symmetric polynomials* are

$$s_r(X_1, \dots, X_n) = \sum_{i_1 < \dots < i_r} X_{i_1} X_{i_2} \cdots X_{i_r}$$

For instance,

$$s_2(X_1, X_2, X_3) = X_1 X_2 + X_1 X_3 + X_2 X_3$$

It is clear that these are symmetric polynomials.

Definition. A *monomial* is an expression of the form $X_I = X_1^{I_1} \cdots X_n^{I_n}$ for $I \in \mathbb{N}^n$. The *(total) degree* of a monomial is $\sum_{i=1}^n I_i$. A *term* is a scalar multiple of a monomial. A polynomial is uniquely characterised by a sum of terms. The total degree of a polynomial is the maximum total degree of its terms.

Monomials are equipped with a *lexicographic ordering*, where we say monomials $X_I > X_J$ if either $I_1 > J_1$ or $I_1 = J_1$ and for some $r \in \{1, \dots, n-1\}$ we have $I_1 = J_1, \dots, I_r = J_r, I_{r+1} > J_{r+1}$. This is a total order.

Theorem. Every symmetric polynomial in n variables over a ring R can be expressed as a polynomial in the s_r for $1 \leq r \leq n$, with coefficients in R. Further, there are no non-trivial relations between the s_r.

Remark. Consider the ring homomorphism $\theta : R[Y_1, \dots, Y_n] \to R[X_1, \dots, X_n]$ given by $\theta(Y_r) = s_r$ and $\theta(r) = r$ for $r \in R$. The first part of the above theorem stipulates that $\operatorname{Im} \theta$ is the set of symmetric polynomials. The second part implies that θ is injective, since any element of $\ker \theta$ is a polynomial between the s_r that evaluates to zero.

Note that we can equivalently define the s_r as

$$\prod_{i=1}^n (T + X_i) = T^n + s_1 T^{n-1} + \dots + s_{n-1} T + s_n$$

If we need to specify the number of variables, we use $s_{r,n}$ instead of s_r.

Proof. Let d be the total degree of a symmetric polynomial f. Let X_I be the largest (in lexicographic order) monomial which occurs in f with coefficient c. Since f is symmetric, any permutation of the X_i yields another monomial that occurs in f. Hence, $I_1 \geq I_2 \geq \dots \geq I_n$, because otherwise the rearranged monomial that satisfies this will be a strictly larger monomial in f. We can therefore write

$$X_I = X_1^{I_1 - I_2} (X_1 X_2)^{I_2 - I_3} \cdots (X_1 \dots X_n)^{I_n}$$

Consider

$$g = s_1^{I_1 - I_2} s_2^{I_2 - I_3} \cdots s_n^{I_n}$$

By construction, the largest monomial in g is X_I. Since g is symmetric, cg is symmetric. By induction, we may assume $f - cg$ is expressible as a sum of symmetric polynomials as it has total degree no larger than d, its leading monomial term is smaller than X_I, and there are only finitely many monomials of degree at most d. Hence f is also expressible as a sum of polynomials as required.

Now we prove uniqueness by induction on n. Let $G \in R[Y_1, \dots, Y_n]$ such that $G(s_{1,n}, \dots, s_{n,n}) = 0$. We want to show that G is the zero polynomial. If $n = 1$, the result is trivial as $s_{1,1} = X_1$. If $G = Y_n^k H$ with Y_n not dividing H, then $s_{n,n}^k H(s_{1,n}, \dots, s_{n,n}) = 0$. Since $s_{n,n} = X_1 \dots X_n$, it is not a zero divisor in $R[X_1, \dots, X_n]$. Hence $H(s_{1,n}, \dots, s_{n,n}) = 0$. Without loss of generality, we can assume that G is not divisible by Y_n. Now, replacing X_n with zero, $s_{k,n}$ is mapped to $s_{k,n-1}$ for $k \neq n$, and $s_{n,n}$ is mapped to zero. Hence, $G(s_{1,n-1}, \dots, s_{n-1,n-1}, 0) = 0$. By induction, $G(Y_1, \dots, Y_{n-1}, 0) = 0$. Hence $Y_n \mid G$, contradicting our assumption. □

Example. Consider, for $n \geq 3$,

$$f = \sum_{i \neq j} X_i^2 X_j$$

The leading term is $X_1^2 X_2 = X_1(X_1 X_2)$, so we consider

$$\begin{aligned} f - s_1 s_2 &= \left(\sum_{i \neq j} X_i^2 X_j \right) - \sum_i \sum_{j<k} X_i X_j X_k \\ &= \left(\sum_{i \neq j} X_i^2 X_j \right) - \left(\sum_{i \neq j} X_i^2 X_j + 3 \sum_{i<j<k} X_i X_j X_k \right) \\ &= -3 \sum_{i<j<k} X_i X_j X_k \\ &= -3 s_3 \end{aligned}$$

Hence $f = s_1 s_2 - 3 s_3$.

Consider $f = p_5 = \sum_i X_i^5$. Computing this in terms of elementary symmetric polynomials by hand is somewhat tedious, but there are various results, such as Newton's formulae, which can help in simplifying such expressions.

Theorem (Newton's formulae)**.** Let $n \geq 1$. Then for all $k \geq 1$,

$$p_k - s_1 p_{k-1} + \dots + (-1)^{k-1} s_{k-1} p_1 + (-1)^k k s_k = 0$$

By convention, let $s_0 = 1$ and $s_r = 0$ if $r > n$.

Proof. It suffices to consider $R = \mathbb{Z}$ (or, for example, $R = \mathbb{R}$). Consider the generating function

$$F(T) = \prod_{i=1}^n (1 - X_i T) = \sum_{r=0}^n (-1)^r s_r T^r$$

Note that for polynomials $f(x), g(x)$, their formal derivatives satisfy

$$\frac{\frac{\mathrm{d}}{\mathrm{d}T}(fg)}{fg} = \frac{f'g + fg'}{fg} = \frac{f'}{f} + \frac{g'}{g}$$

Then, taking the logarithmic derivative with respect to T,

$$\begin{aligned}\frac{F'(T)}{F(T)} &= \frac{\frac{\mathrm{d}}{\mathrm{d}T}\prod_{i=1}^n(1-X_iT)}{\prod_{i=1}^n(1-X_iT)} \\ &= \sum_{i=1}^n \frac{\frac{\mathrm{d}}{\mathrm{d}T}(1-X_iT)}{1-X_iT} \\ &= -\sum_{i=1}^n \frac{X_i}{1-X_iT} \\ &= \frac{-1}{T}\sum_{i=1}^n\sum_{r=1}^\infty X_i^rT^r \\ &= \frac{-1}{T}\sum_{r=1}^\infty p_rT^r\end{aligned}$$

Hence,

$$-TF'(T) = s_1T - 2s_2T^2 + \dots + (-1)^{n-1}ns_nT^n$$

but also

$$-TF'(T) = F(T)\sum_{r=1}^\infty p_rT_r = (s_0 - s_1T + \dots + (-1)^n s_nT^n)\big(p_1T + p_2T^2 + \dots\big)$$

Equating the coefficients of powers of T, we find the identity as required by the theorem. □

Example. The *discriminant polynomial* is

$$D(X_1,\dots,X_n) = \Delta(X_1,\dots,X_n)^2$$

where

$$\Delta(X_1,\dots,X_n) = \prod_{i<j}(X_i - X_j)$$

This is used in defining the sign of a permutation: applying a permutation σ to Δ multiplies Δ by the sign of σ. Hence D is symmetric. Therefore, D can be written in terms of the symmetric polynomials.

$$D(X_1,\dots,X_n) = d(s_1,\dots,s_n)$$

where d has integer coefficients. For example, $n = 2$ gives $D = (X_1 - X_2)^2 = s_1^2 - 4s_2$.

Definition. Let $f = T^n + \sum_{i=0}^{n-1} a_{n-i}T^i$ be a monic polynomial in $R[T]$. Its discriminant is

$$\mathrm{Disc}(f) = d(-a_1, a_2, -a_3, \dots, (-1)^n a_n) \in R$$

Observe that if f is a product of linear polynomials $f = \prod_{i=1}^{n}(T - x_i)$, then

$$a_r = (-1)^r s_r(x_1, \dots, x_n)$$

giving

$$\operatorname{Disc}(f) = \prod_{i<j}(x_i - x_j)^2 = D(x_1, \dots, x_n)$$

In particular, if $R = K$ is a field, $\operatorname{Disc}(f) = 0$ if and only if f has a repeated root. For example, $\operatorname{Disc}(T^2 + bT + c) = b^2 - 4c$.

2. Fields

2.1. Definition

Definition. A *field* is a commutative nonzero ring K with a 1, in which every nonzero element is invertible. The set of nonzero elements in K is therefore a group under multiplication, known as the multiplicative group of K, denoted $K^\times$.

Definition. The *characteristic* of a field K is the least positive integer p such that $p \cdot 1 = 0$; or if such an integer does not exist, its characteristic is zero.

Example. $\mathbb{Q}$ has characteristic zero. $\mathbb{F}_p = {}^{\mathbb{Z}}\!/_{p\mathbb{Z}}$ has characteristic p, when p is prime.

Remark. The characteristic of a field is always prime or zero.

Definition. The *prime subfield* of a field K is the smallest subfield of K, which is isomorphic to $\mathbb{F}_p$ (if its characteristic is a prime p) or $\mathbb{Q}$ (if its characteristic is zero).

Proposition. Let $\varphi : K \to L$ be a field homomorphism. Then φ is an injection.

Proof. We have $\varphi(1_K) = 1_L \neq 0_L$ by the definition of a ring homomorphism. Then $\ker\varphi$ is a proper ideal of K. But the only proper ideal of a field is the zero ideal, so $\ker\varphi = (0)$. □

2.2. Field extensions

Definition. Let $K \subset L$ be fields (implicitly assuming that the field operations and identity elements on K and L are the same). We say K is a subfield of L, and L is a *field extension* of K, denoted L/K (read 'L over K'). If $i : K \to L$ is a field homomorphism, we say that i is an isomorphism of K with the subfield $i(K) \subset L$; in this case, we identify K with $i(K)$ and say L is a field extension of K.

Remark. The notation L/K is not related to quotients or division.

Example. (i) $\mathbb{C}/\mathbb{R}/\mathbb{Q}$.

(ii) $\mathbb{Q}(i) = \{a + bi \mid a, b \in \mathbb{Q}\}/\mathbb{Q}$.

Definition. Let $K \subset L$, and $x \in L$. We define $K[x] = \{p(x) \mid p \in K[T]\}$, the ring of polynomial expressions in x. This is a subring of L, but is not in general a field. We further define $K(x) = \left\{\frac{p(x)}{q(x)} \;\middle|\; p, q \in K[T], q(x) \neq 0\right\}$ to be the field of fractions of $K[x]$, which is the field of rational expressions in x. This is a subfield of L, usually read 'K adjoin x'. For $x_1, \dots, x_n \in L$, we define

$$K[x_1, \dots, x_n] = \{p(x_1, \dots, x_n) \mid p \in K[T_1, \dots, T_n]\}$$
$$K(x_1, \dots, x_n) = \left\{\frac{p(x_1, \dots, x_n)}{q(x_1, \dots, x_n)} \;\middle|\; p, q \in K[T_1, \dots, T_n], q(x_1, \dots, x_n) \neq 0\right\}$$

Remark. One can check that $K(x_1, \dots, x_{n-1})(x_n) = K(x_1, \dots, x_n)$ and similarly for $K[x_1, \dots, x_n]$.

2.3. Field extensions as vector spaces

Remark. A field extension L/K turns L into a K-vector space by forgetting the multiplication between elements of L.

Definition. A field extension L/K is called a *finite extension* if L is a finite-dimensional K-vector space. In this case, we write $[L : K] = \dim_K L$ for the dimension of this vector space, known as the *degree* of the extension. Otherwise, we say L/K is an *infinite extension*, and write $[L : K] = \infty$.

Remark. $[L : L] = \dim_L L = 1$. As a K-vector space, $L \cong K^{[L:K]}$.

Example. $\mathbb{C}/\mathbb{R}$ is a finite extension of degree two.

If K is any field, the extension $K(X)/K$ is an infinite extension, where $K(X)$ is the field of rational functions, the field of fractions of the polynomial ring $K[X]$. This is because $1, X, X^2, \dots$ are linearly independent.

$\mathbb{R}/\mathbb{Q}$ is an infinite extension. This follows by a countability argument. If $\mathbb{R}/\mathbb{Q}$ were a finite extension of degree n, we would have $\mathbb{R} \cong \mathbb{Q}^n$, but the left hand side is uncountable and the right hand side is countable.

This course is largely concerned with properties and symmetries of finite field extensions.

Definition. An extension is *quadratic*, *cubic*, etc. if its degree is 2, 3, etc.

Proposition. Suppose K is a finite field (necessarily of characteristic p for $p \neq 0$ a prime). Then $|K|$ is a power of p.

Proof. Note that $K/\mathbb{F}_p$ is a finite extension, and so $K \cong \mathbb{F}_p^n$, giving $|K| = p^n$. □

We will later show that for all prime powers $q = p^n$, there exists a finite field $\mathbb{F}_q$ with q elements.

Theorem (tower law)**.** Let M/L, L/K be a pair of field extensions. Then M/K is a finite extension if and only if M/L and L/K are finite. If so, we have $[M : L][L : K] = [M : K]$.

It is easier to prove a more general statement.

Theorem. Let L/K and V is an L-vector space. Then V is a K-vector space, and $\dim_K V = [L : K] \dim_L V$ (with the obvious meaning if any of these expressions are infinite).

Taking $V = M$ proves the tower law as required.

Proof. Let $\dim_L V = d < \infty$. Then $V \cong L \oplus \dots \oplus L = L^d$ as an L-vector space, so this also holds as a K-vector space. But since $L \cong K^{[L:K]}$ as a K-vector space, we have $V \cong (K^{[L:K]})^d \cong K^{d[L:K]}$ as a K-vector space.

If V is finite-dimensional over K, then a K-basis for V will span V over L, so V is finite-dimensional over L. Thus if V is infinite-dimensional over L, it is infinite-dimensional over K.

Likewise, if $[L : K] = \infty$ and $V \neq 0$, then V has an infinite set of linearly independent vectors as a K-vector space, so $\dim_K V = \infty$. □

Proposition. Let K be a field, and $G \subset K^\times$ be a finite subgroup of the multiplicative group. Then G is cyclic. In particular, if K is finite, $K^\times$ is cyclic.

Proof. We can find m_i such that

$$G \cong \mathbb{Z}/_{m_1\mathbb{Z}} \times \dots \times \mathbb{Z}/_{m_k\mathbb{Z}}$$

where $1 < m_1 \mid m_2 \mid \dots \mid m_k = m$ by the structure theorem for abelian groups. Then, every element of G satisfies $x^m = 1$. Since K is a field, the polynomial $T^m - 1$ has at most m roots. Every element of G is a root of this polynomial, so $|G| \leq m$. This can only happen when $k = 1$, so $G = \mathbb{Z}/_{m\mathbb{Z}}$. □

Remark. If $K = \mathbb{F}_p = \mathbb{Z}/_{p\mathbb{Z}}$, there exists $a \in \{1, \dots, p-1\}$ such that $\mathbb{Z}/_{p\mathbb{Z}} = \{1, a, a^2, \dots, a^{p-1}\}$. Such an a is called a *primitive root* mod p.

Proposition. Let R be a ring, p be a prime such that $p \cdot 1_R = 0_R$ (for instance, R could be a field of characteristic p). Then, the map $\varphi_p : R \to R$ given by $\varphi_p(x) = x^p$ is a homomorphism, known as the *Frobenius endomorphism*.

Proof. First, $\varphi_p(1) = 1^p = 1$ and $\varphi_p(x)\varphi_p(y) = x^p y^p = (xy)^p = \varphi_p(xy)$. For $x, y \in R$,

$$\begin{aligned}\varphi_p(x+y) &= \binom{p}{0}x^p y^0 + \binom{p}{1}x^{p-1}y^1 + \dots + \binom{p}{p-1}x^1 y^{p-1} + \binom{p}{p}x^0 y^p \\ &= x^p + y^p = \varphi_p(x) + \varphi_p(y)\end{aligned}$$

since $p \mid \binom{p}{k}$ for $k \in \{1, \dots, p-1\}$ by primality of p. □

Example. This gives another proof of Fermat's little theorem $x^p \equiv x \bmod p$, by induction on x noting that $(x+1)^p \equiv x^p + 1 \bmod p$.

2.4. Algebraic elements and minimal polynomials

Definition. Let L/K be an extension and $x \in L$. x is *algebraic over K* if there exists a nonzero polynomial $f \in K[T]$ such that $f(x) = 0$. Otherwise, we say x is *transcendental over K*.

For $f \in K[T]$, we have $f(x) \in L$. Varying f, this gives a map $\mathrm{ev}_x : K[T] \to L$ defined by $f \mapsto f(x)$. This is a ring homomorphism.

The kernel $I = \ker(\mathrm{ev}_x) \subset K[T]$ is an ideal, the set of polynomials which vanish at x. As $\mathrm{Im}(\mathrm{ev}_x)$ is a subring of L which is a field, it is an integral domain. In particular, I is a prime ideal, so either $I = 0$, in which case x is transcendental over K, or there exists a unique monic irreducible polynomial $0 \neq g \in K[T]$ such that $I = (g)$, in which case x is algebraic and we say g is the *minimal polynomial* of x over K. In this case, $f(x) = 0$ if and only if $g \mid f$. We write $m_{x,K}$ for the minimal polynomial of x over K. Note that $m_{x,K}$ is the monic polynomial in K of least degree with x as a root.

Example. If $x \in K$, $m_{x,K} = T - x$. If p is prime and $d \geq 1$, $T^d - p \in \mathbb{Q}[T]$ is irreducible by Eisenstein's criterion, so it is the minimal polynomial of $\sqrt[d]{p} \in \mathbb{R}$ over $\mathbb{Q}$. If p is prime, $z = e^{\frac{2\pi i}{p}}$ is a root of $T^p - 1 = (T-1)(T^{p-1} + T^{p-2} + \dots + 1) = (T-1)g(T)$. Note that

$$g(T+1) = \binom{p}{p}T^{p-1} + \binom{p}{p-1}T^{p-2} + \dots + \binom{p}{2}T + \binom{p}{1}$$

This is irreducible by Eisenstein's criterion, so g is minimal for z over $\mathbb{Q}$.

We say the degree of an algebraic element x over K is the degree of its minimal polynomial, written $\deg_K x = \deg(x/K)$.

Proposition. Let L/K and $x \in L$. Then, the following are equivalent.

(i) x is algebraic over K.

(ii) $[K(x) : K]$ is finite.

(iii) $K[x]$ is finite-dimensional as a K-vector space.

(iv) $K[x] = K(x)$.

(v) $K[x]$ is a field.

If these hold, $\deg x = [K(x) : K]$.

Proof. (ii) implies (iii). This follows since $K[x] \subseteq K(x)$.

(iv) is equivalent to (v) is trivial.

(iii) implies (v) and (ii). Let $0 \neq y = g(x) \in K[x]$. Consider the map $K[x] \to K[x]$ given by $z \mapsto yz$. This is a K-linear transformation, and since $y \neq 0$ this is injective. Because $\dim K[x]$ is finite, this injective map must be a bijection. Therefore there exists z such that $yz = 1$, so y is invertible. Hence (v) holds. Since (v) implies (iv), $[K(x) : K] = \dim_K K[x] < \infty$ as required for (ii).

(v) implies (i). If $x = 0$, the proof is complete, so assume $x \neq 0$. Then $x^{-1} = a_0 + a_1 x + \dots + a_n x^n \in K[x]$. Therefore, $a_n x^{n+1} + \dots + a_0 x - 1 = 0$, so x is algebraic over K.

(i) implies (v), (iii), and the degree formula. The image of $\mathrm{ev}_x : K[T] \to L$ is the subring $K[x] \subset L$. If x is algebraic over K, $\ker(\mathrm{ev}_x) = (m_{x,K})$ is a maximal ideal by irreducibility of $m_{x,K}$. By the first isomorphism theorem, ${}^{K[T]}\!/\!_{(m_{x,K})} \cong K[x]$. But quotients by maximal ideals are fields, so $K[x]$ is a field, proving (v). This polynomial is monic of degree $d = \deg_K x$. Hence ${}^{K[T]}\!/\!_{(m_{x,K})}$ has a K-basis $1, T, \dots, T^{d-1}$. Thus, $\dim_K K[x] = d = [K(x) : K] < \infty$, proving (iii) and the degree formula. □

Corollary. $x_1, \dots, x_n$ are algebraic over K if and only if $L = K(x_1, \dots, x_n)$ is finite over K. If so, every element of $K(x_1, \dots, x_n)$ is algebraic over K.

If x, y are algebraic over K, then so are $x \pm y$, xy, and x^{-1} if x is nonzero. If L/K is a field extension, the set of algebraic elements of L forms a subfield of L.

Proof. If x_n is algebraic over K, then it is also algebraic over $K(x_1, \dots, x_{n-1})$. Hence the extension $L/K(x_1, \dots, x_{n-1})$ is finite. By induction on n, the tower law gives the required result. Conversely, if L is finite over K, the subfield $K(y)$ is finite over K for all $y \in L$, so y is algebraic over K.

Suppose x, y are algebraic over K. Then $x \pm y, xy, x^{-1} \in K(x, y)$, which is finite over K as required. □

Example. Consider $z = e^{2\pi i/p} \in \mathbb{C}$ where p is an odd prime. This has degree $p - 1$ as discussed above. Now consider $x = 2\cos\frac{2\pi}{p}$, so $x = z + \frac{1}{z} \in \mathbb{Q}(z)$. This is algebraic over $\mathbb{Q}$ because it belongs to this finite extension. Note that $\mathbb{Q}(z) \supset \mathbb{Q}(x) \supset \mathbb{Q}$, and $z^2 - xz + 1 = 0$. Hence the degree of z over $\mathbb{Q}(x)$ is at most 2. But $[\mathbb{Q}(z) : \mathbb{Q}(x)] \neq 1$ because $z \in \mathbb{C} \setminus \mathbb{R}$. By the tower law, we must have $[\mathbb{Q}(z) : \mathbb{Q}] = \frac{p-1}{2}$.

We can now derive the minimal polynomial by considering $z^{\frac{p-1}{2}} + z^{\frac{p-3}{2}} + \dots + z^{\frac{-p-1}{2}} = 0$. Since $z + z^{-1} = x$, we can express this as a polynomial in x of degree $\frac{p-1}{2}$.

Example. Let $x = \sqrt{m} + \sqrt{n}$ where m, n are integers, and m, n, mn are not squares. We know that $n = (x - \sqrt{m})^2 = x^2 - 2x\sqrt{m} + m$, so $[\mathbb{Q}(x) : \mathbb{Q}(\sqrt{m})] \leq 2$. By symmetry, $[\mathbb{Q}(x) : \mathbb{Q}(\sqrt{n})] \leq 2$. Note that $\sqrt{m} \in \mathbb{Q}(x)$ because $\frac{x^2+m-n}{2x} = \sqrt{m}$.

m, n are not squares, so $[\mathbb{Q}(\sqrt{m}) : \mathbb{Q}] = 2$. By the tower law we have $[\mathbb{Q}(x) : \mathbb{Q}] \in \{2, 4\}$. If $[\mathbb{Q}(x) : \mathbb{Q}] = 2$, we have $\mathbb{Q}(x) = \mathbb{Q}(\sqrt{m}) = \mathbb{Q}(\sqrt{n})$. In this case, $\sqrt{m} = a + b\sqrt{n} \implies m = a^2 + b^2 n + 2ab\sqrt{n}$, but n is not a square, so by rationality, $ab = 0$. But if $b = 0$, m is a square, and if $a = 0$, $mn = b^2 n^2$ is a square. Hence the degree of the field extension is 4.

Definition. An extension L/K is *algebraic* if all elements of L are algebraic over K.

Lemma. Let $M/L/K$, where L/K is algebraic. Suppose x is algebraic over L. Then x is algebraic over K.

Proof. There exists $f = T^n + a_{n-1}T^{n-1} + \dots + a_0 \in L[T]$ where $f \neq 0$ and $f(x) = 0$. Let $L_0 = K(a_0, \dots, a_{n-1})$. As each $a_i \in L$ is algebraic over K, we must have that $[L_0 : K]$ is finite. As $f \in L_0[T]$, x is algebraic over L_0. So $[L_0(x) : L_0] < \infty \implies [L_0(x) : K] < \infty$. Hence $[K(x) : K] < \infty$, so x is algebraic over K. □

Proposition. (i) Finite extensions are algebraic.

(ii) $K(x)$ is algebraic over K if and only if x is algebraic over K.

(iii) If $M/L/K$, we have M/K is algebraic if and only if M/L and L/K are algebraic.

Proof. (i) $[L : K] < \infty$, so for all $x \in L$, $[K(x) : K] < \infty$, so x is algebraic.

(ii) Certainly if $K(x)$ is algebraic over K, we have that x is algebraic over K. Conversely, if x is algebraic over K, $[K(x) : K]$ is finite, so it is algebraic by part (i).

(iii) Suppose M/K is algebraic. Then for all $x \in M$, we have that x is algebraic over K, so it satisfies a polynomial $f \in K[T]$. Hence $f \in L[T]$ is another polynomial that x satisfies, so M/L is algebraic. L/K is clearly algebraic because it is contained within M.

Conversely, suppose M/L and L/K are algebraic. Let $x \in M$. Then by the previous lemma, x is algebraic over K as required.

□

Example. Let $K = \mathbb{Q}$ and $L = \{x \in \mathbb{C} \mid x \text{ is algebraic over } \mathbb{Q}\} = \overline{\mathbb{Q}}$. This extension $\overline{\mathbb{Q}}/\mathbb{Q}$ is algebraic, but not finite. Indeed, for every $n \geq 1$, $\sqrt[n]{2} \in L$, and $[\mathbb{Q}(\sqrt[n]{2}) : \mathbb{Q}] = n$ by irreducibility of $T^n - 2$. In particular, L contains subfields of arbitrarily large degree, so cannot be a finite extension.

2.5. Algebraic numbers in the real line and complex plane

Traditionally, we call $x \in \mathbb{C}$ algebraic if it is algebraic over $\mathbb{Q}$, otherwise it is transcendental. $\overline{\mathbb{Q}} = \{x \mid x \text{ algebraic}\}$ is a proper subfield of $\mathbb{C}$. Indeed, $\mathbb{Q}[T]$ is a countable set, and $\mathbb{C}$ is uncountable. However, it is difficult to explicitly find an element of $\mathbb{C} \setminus \overline{\mathbb{Q}}$, or to show that a given number is transcendental.

Example. Liouville's constant $c = \sum_{n\geq 1} 10^{-n!}$ is transcendental, as proven in IA Numbers and Sets. This can be proven by showing that algebraic numbers cannot be 'well approximated' by rationals.

Example. Hermite and Lindemann showed that e and π are transcendental.

Example. Let x, y be algebraic, and $x \neq 0, 1$. Gelfond and Schneider showed that x^y is algebraic if and only if y is rational. In particular, $e^\pi = (-1)^{-i}$ is transcendental.

2.6. Ruler and compass constructions

Definition. A *ruler and compass construction* in plane geometry is a drawing constructed with the following methods.

(i) Given P_1, P_2, Q_1, Q_2 in the plane and $P_i \neq Q_i$, we can construct the point of intersection of the lines P_1Q_1 and P_2Q_2, if indeed they do intersect.

(ii) Given P_1, P_2, Q_1, Q_2 in the plane and $P_i \neq Q_i$, we can construct the points of intersection of the circles with centres P_i that pass through the Q_i, if they intersect.

(iii) Similarly we can construct the points of intersection of a line and a circle.

A point $(x, y) \in \mathbb{R}^2$ is *constructible* from a set $\{(x_1, y_1), \dots, (x_n, y_n)\}$ if it can be obtained by finitely many expansions of the set under applications of the above operations. A real number $x \in \mathbb{R}$ is *constructible* if $(x, 0)$ is constructible from $\{(0, 0), (1, 0)\}$.

Remark. Every rational is constructible. Square roots of constructible numbers are constructible.

Definition. Let $K \subseteq \mathbb{R}$ be a subfield of the reals. We say K is *constructible* if there exists $n \in \mathbb{N}$ and fields $\mathbb{Q} = F_0 \subset F_1 \subset \dots \subset F_n \subseteq \mathbb{R}$ and $a_i \in F_i$ for $1 \leq i \leq n$ such that

(i) $K \subseteq F_n$;

(ii) $F_i = F_{i-1}(a_i)$;

(iii) $a_i^2 \in F_{i-1}$.

Remark. By conditions (ii) and (iii), F_i/F_{i-1} is at most a quadratic extension. Then, by the tower law, $F_n/\mathbb{Q}$ has degree a power of two, so $K/\mathbb{Q}$ is a finite extension with degree a power of two.

Theorem. If x is constructible, $\mathbb{Q}(x)$ is constructible.

Proof. Let $K = \mathbb{Q}(x)$. We show that if (x, y) can be constructed with k steps, $\mathbb{Q}(x, y)$ is a constructible extension of $\mathbb{Q}$. By induction, suppose $\mathbb{Q} = F_0 \subset \dots \subset F_n$ satisfy conditions (ii) and (iii) such that the coordinates of the points obtained after $k - 1$ constructions lie in F_n.

The intersection point of two lines has coordinates given by rational functions of the coordinates of the points P_i, Q_i with rational coefficients. In particular, if the kth construction is of this type, the intersection point has coordinates in F_n. We can similarly see that the intersection points of two circles and the intersection points of a line and a circle have coordinates given by quadratic equations $a \pm b\sqrt{e}, c \pm d\sqrt{e}$, where a, b, c, d, e are rational functions of the coordinates P_i, Q_i. Thus the new points have coordinates which lie in $F_n(\sqrt{e})$, a constructible extension of $\mathbb{Q}$ as required. □

Corollary. If x is constructible, x is algebraic over $\mathbb{Q}$ and the degree of the minimal polynomial is a power of two.

Remark. One can show that if $\mathbb{Q}(x)$ is constructible, we also have x is constructible, so the above theorem is a bi-implication. However, this will not be required for our purposes in this course.

2.7. Classical problems

Theorem. It is impossible to square the circle.

Proof. The statement is to construct a square with area equal to that of a given circle. In particular, we must construct $\sqrt{\pi}$. Suppose such a construction can occur. Then π is also constructible. But π is transcendental and hence inconstructible. □

Theorem. It is impossible to duplicate the cube.

Proof. To duplicate the cube, one must be able to construct $\sqrt[3]{2}$. The minimal polynomial of $\sqrt[3]{2}$ is $X^3 - 2$. This can be easily checked with Eisenstein's criterion. Since the minimal polynomial is of degree not a power of two, $\sqrt[3]{2}$ is inconstructible. □

Theorem. It is impossible to trisect a given angle.

Proof. If we can trisect any constructible angle, we can in particular trisect the (constructible) angle $\frac{2\pi}{3}$, for example to construct a regular nonagon. Then the angle $\frac{2\pi}{9}$ would be constructible, so its sine and cosine would be constructible. By the triple angle formula for cosine,

$$\cos 3\theta = 4\cos^3\theta - 3\cos\theta \implies 4\cos\left(\frac{2\pi}{9}\right)^3 - 3\cos\left(\frac{2\pi}{9}\right) = \cos\left(\frac{2\pi}{3}\right)$$

Hence $\cos\left(\frac{2\pi}{9}\right)$ is a root of $8X^3 - 6X + 1$. In particular, $2\cos\left(\frac{2\pi}{9}\right) - 2$ is a root of $X^3 + 6X^2 + 9X + 3$, which can be shown to be irreducible by Eisenstein's criterion. But this has degree 3, so $\deg_{\mathbb{Q}} \cos\left(\frac{2\pi}{9}\right) = 3$, so this is inconstructible. In particular, the regular nonagon is inconstructible. □

We will later prove the following theorem.

Theorem (Gauss)**.** A regular n-gon is is constructible if and only if n is the product of a power of two and distinct *Fermat primes*, which are the primes of the form $2^{2^k} + 1$.

3. Types of field extensions

3.1. Fields from polynomials

Suppose K is a field and $f \in K[T]$. We wish to find an extension L/K of degree as small as possible such that f is expressible as a product of linear factors in $L[T]$.

Example. Let $K = \mathbb{Q}$. Then by the fundamental theorem of algebra, a monic polynomial $f \in \mathbb{Q}[T]$ is expressible as a product of n linear factors $(T - x_i)$ in $\mathbb{C}[T]$. One example of such a field extension is $L = \mathbb{Q}(x_1, \dots, x_n)$, which is a finite extension of $\mathbb{Q}$.

We will later give another proof of the fundamental theorem of algebra using techniques from Galois theory.

Example. Let $K = \mathbb{F}_p$, and f is irreducible and has degree $d > 1$. Since there is no ambient field structure, explicitly finding L is more challenging. We will first find an extension in which f has at least one root, and then use induction.

Theorem. Let f be a monic irreducible polynomial. Let $L_f = {}^{K[T]}\!/\!_{(f)}$. Since f is irreducible, (f) is maximal, hence L_f is a field. Let $t \in L_f$ be the residue class T modulo (f). Then L_f/K is a finite field extension of degree $\deg f$, and f is the minimal polynomial for t.

We have thus constructed a field extension of K for which f has at least a single root. Recall that if x is algebraic over K, then $K(x) \cong {}^{K[T]}\!/\!_{(f)}$ where f is minimal for x.

Definition. Let K be a field, and $L/K, M/K$ are field extensions. A *K-homomorphism* or *K-embedding* from L to M is a field homomorphism $\sigma : L \to M$ such that $\sigma|_K = \mathrm{id}_K$.

The naming 'K-embedding' is justified because any field homomorphism is injective.

Theorem. Let $f \in K[T]$ be irreducible, and L/K a field extension. Then:

(i) If $x \in L$ is a root of f, there exists a unique K-homomorphism $\sigma : L_f = {}^{K[T]}\!/\!_{(f)} \to L$ such that $t = T + (f) \mapsto x$.

(ii) Every K-homomorphism $\sigma : L_f \to L$ arises in this way.

Hence, we have a bijection between K-homomorphisms $\sigma : L_f \to L$ and the set of roots of f in L. In particular, there are at most $\deg f$-many K-homomorphisms.

Proof. Let $x \in L$ be a root of f. We define the K-homomorphism $\sigma : {}^{K[T]}\!/\!_{(f)} \to L$ by $\sigma(T) = x$. Conversely, suppose $\sigma : {}^{K[T]}\!/\!_{(f)} \to L$ is a K-homomorphism. Then $\sigma(T)$ is a root of f, because $f(\sigma(T)) = \sigma(f(T)) = \sigma(0) = 0$. So the two definitions are inverses, so we have a one-to-one correspondence as required. □

Corollary. Let $L = K(x)$ for some x algebraic over K. Then there exists a unique isomorphism $\sigma : L_f \to K(x)$ such that $\sigma(t) = x$, where f is minimal for x over K.

Definition. Let x, y be algebraic over K. We say x, y are *K-conjugate* if they have the same minimal polynomial over K.

By the corollary above, $K(x)$ and $K(y)$ are isomorphic to L_f where f is minimal for x and y over K.

Corollary. Algebraic elements x, y are K-conjugate if and only if there exists a K-isomorphism $\sigma : K(x) \to K(y)$ such that $\sigma(x) = y$.

Proof. The above corollary shows the forward direction. Conversely, for all $g \in K[T]$, we have $\sigma(g(x)) = g(\sigma(x))$ so they have the same minimal polynomial. □

Informally, the roots of an irreducible polynomial are algebraically indistinguishable.

It can be useful for inductive arguments to have a generalisation of the above theorem.

Definition. Let $L/K, L'/K'$ be field extensions, and let $\sigma : K \to K'$ be a field homomorphism. Let $\tau : L \to L'$ be a field homomorphism such that $\tau(x) = \sigma(x)$ for all $x \in K$. Then we say τ is a *σ-homomorphism* from L to L'. We also say τ *extends* σ, or that σ is the *restriction* of τ to K.

We can now define the following variant of the previous theorem.

Theorem. Let $f \in K[T]$ be irreducible, and $\sigma : K \to L$ be a field homomorphism. Let σf be the polynomial obtained by applying σ to the coefficients of f.

(i) If $x \in L$ is a root of f, there exists a unique σ-homomorphism $\tau : L_f \to L$ such that $\tau(t) = x$.

(ii) Every σ-homomorphism $L_f \to L$ is of this form.

Therefore there is a bijection between the σ-homomorphisms $L_f \to L$ and the roots of f in L.

Example. Let $K = \mathbb{Q}(\sqrt{2}) \subset \mathbb{R}$, and $L = \mathbb{C}$. Let $\sigma : K \to L$ be the homomorphism such that $\sigma(x + y\sqrt{2}) = x - y\sqrt{2}$. Then let $f = T^2 - (1 + \sqrt{2})$. Then the map $\tau : L_f \to \mathbb{C}$ must satisfy $\tau(t) = \pm\sqrt{1 - \sqrt{2}} = \pm i\sqrt{\sqrt{2} - 1} \in \mathbb{C}$. If instead we let $\sigma(x + y\sqrt{2}) = x + y\sqrt{2}$, we have $\tau(t) = \pm\sqrt{\sqrt{2} + 1}$, which are both real.

3.2. Splitting fields

Definition. Let $f \in K[T]$ be a nonzero polynomial that is not necessarily irreducible. We say that an extension L/K is a *splitting field* for f over K if

(i) f splits into linear factors in $L[T]$;

(ii) $L = K(x_1, \dots, x_n)$, where the x_i are the roots of f in L.

Remark. The second criterion ensures that f does not split into linear factors in any proper subfield of L. Note that any splitting field is finite, because the adjoined elements are algebraic.

Theorem. Every nonzero polynomial has a splitting field.

Proof. Let $f \in K[T]$. We prove this by induction on the degree of f, but allow K to vary. If f is constant, there is nothing to prove, since K is already a splitting field. Suppose that for all fields K' and all polynomials in $K'[T]$ of degree less than f, there is a splitting field. Consider an irreducible factor g of f, and consider $K' = L_g = {}^{K[T]}\!/_{(g)}$. Let $x_1 = T + (g)$. Then $g(x_1) = 0$, so $f(x_1) = 0$, hence $f = (T - x_1)f_1$, where $f_1 \in K'[T]$. By induction, there exists a splitting field L for f_1 over K' since $\deg f_1 < \deg f$. Let $x_2, \dots, x_n \in L$ be the roots of f_1 in L. Then f splits into linear factors in L with roots $\{x_1, x_2, \dots, x_n\}$. Because L is a splitting field for f_1 over K', we have $L = K'(x_2, \dots, x_n) = K(x_1)(x_2, \dots, x_n) = K(x_1, \dots, x_n)$, so L is a splitting field for f. □

Remark. If $K \subseteq \mathbb{C}$, we already know by the fundamental theorem of algebra that any polynomial over K has a subfield of $\mathbb{C}$ as its splitting field.

Theorem. Let $f \in K[T]$ be a polynomial and L/K be a splitting field for f. Then let $\sigma : K \to M$ be a field homomorphism such that σf splits in $M[T]$. Then

(i) σ can be extended to a homomorphism $\tau : L \to M$;

(ii) if M is a splitting field for σf over σK, then any $\tau : L \to M$ is an isomorphism.

In particular, any two splitting fields are K-isomorphic.

Remark. When constructing the splitting field for a polynomial, we had choice in which irreducible factors to consider first. It is not clear, without this theorem, that two splitting fields have the same degree.

Note that we can have different $\tau_1, \tau_2 : L \to M$ for splitting fields L, M of f over K.

Proof. We will prove (i) by induction on $[L : K]$. If $n = 1$, we have $L = K$ and there is nothing to prove. Suppose $x \in L \setminus K$ is a root of an irreducible factor g of f in K, so $\deg g > 1$. Let $y \in M$ be a root of $\sigma g \in M[T]$, which exists because σf splits in M. Then, there exists $\sigma_1 : K(x) \to M$ such that $\sigma_1(x) = y$, and σ_1 extends σ. Then, $[L : K(x)] < [L : K]$, so by induction, $\sigma_1 : K(x) \to M$ can be extended to $\tau : L \to M$, because L is a splitting field for f over $K(x)$. This τ therefore extends σ as required.

To prove (ii), suppose M is a splitting field for σf over σK. Let τ be as in (i), and $\{x_i\}$ be the roots of f in L. Then the roots of σf in M are $\{\tau(x_i)\}$. Since M is a splitting field, $M = \sigma K(\{\tau(x_i)\}) = \tau L$ as $L = K(\{x_i\})$. So τ is an isomorphism.

If $K \subseteq M$ and σ is the inclusion homomorphism, τ is a K-isomorphism. □

Example. Let $f = T^3 - 2 \in \mathbb{Q}[T]$. This has splitting field $L = \mathbb{Q}(\sqrt[3]{2}, \omega) \subseteq \mathbb{C}$ where $\omega = e^{\frac{2\pi i}{3}}$. We know $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3$, but $\omega \notin \mathbb{R}$ and $\omega^2 + \omega + 1 = 0$, so $[L : \mathbb{Q}(\sqrt[2]{3})] = 2$ giving $[L : \mathbb{Q}] = 6$ by the tower law. In particular, adjoining a single root to $\mathbb{Q}$ is not enough to generate L.

Example. Let $f = \frac{T^5-1}{T-1} = T^4 + \dots + T + 1 \in \mathbb{Q}[T]$. Let $z = e^{\frac{2\pi i}{5}}$, then this is the minimal polynomial of z. We find $f = \prod_{1 \le a \le 4}(T - z^a)$, so $\mathbb{Q}(z)$ is already a splitting field for f over $\mathbb{Q}$, and $[\mathbb{Q}(z) : \mathbb{Q}] = 4$.

Example. Let $f = T^3 - 2 \in \mathbb{F}_7[T]$. This is irreducible because 2 is not a cube in $\mathbb{F}_7$. Consider $L = {}^{\mathbb{F}_7[X]}\!/\!_{X^3 - 2} = \mathbb{F}_7(x)$, so $x^3 = 2$. Since $2^3 = 4^3 = 1$ in $\mathbb{F}_7$, we have $(2x)^3 = (4x)^3 = 2$, so $x, 2x, 4x$ are roots of f in L. In particular, L is a splitting field for f, since $f = (T - x)(T - 2x)(T - 4x)$; here, adjoining one root is enough to make f split.

3.3. Normal extensions

Definition. An extension L/K is a *normal extension* if it is algebraic and for all $x \in L$, the minimal polynomial splits in L.

Remark. This condition is equivalent to the statement that for every $x \in L$, L contains a splitting field for x. In other words, if an irreducible polynomial $f \in K[T]$ has a single root in L, it splits and has all roots in L.

Theorem. Let L/K be a finite extension. Then L is normal over K if and only if L is a splitting field for some (not necessarily irreducible) polynomial $f \in K[T]$.

Proof. Suppose L is normal. Then $L = K(x_1, \dots, x_n)$ since L is algebraic. Then the minimal polynomial $m_{x_i,K}$ of each x_i over K splits in L. L is generated by the roots of $\prod_i m_{x_i,K}$, so L is a splitting field for f.

For the converse, suppose L is a splitting field for $f \in K[T]$. Let $x \in L$, and let $g = m_{x,K}$ be its minimal polynomial. We want to show that g splits in L. Let M be a splitting field for g over L, and let $y \in M$ be a root of g. We want to show $y \in L$.

Since L is a splitting field for f over K, L is a splitting field for f over $K(x)$, and $L(y)$ is a splitting field for f over $K(y)$. Now, there exists a K-isomorphism between $K(x)$ and $K(y)$, because x, y are roots of the same irreducible polynomial g. By the uniqueness of splitting fields, $[L : K(x)] = [L(y) : K(y)]$. Multiplying by $[K(x) : K]$, we find $[L : K] = [L(y) : K]$ because $[K(y) : K] = [K(x) : K]$ as they are roots of the same irreducible polynomial. Hence $[L(y) : L] = 1$, so $y \in L$ as required. □

Corollary (normal closure). Let L/K be a finite extension. Then there exists a finite extension M/L such that M/K is normal, and if $L \subseteq M' \subseteq M$ and M'/K is normal, $M = M'$. Moreover, any two such extensions M are L-isomorphic.

Such an M is said to be a *normal closure* of L/K.

Proof. Let $L = K(x_1, \dots, x_k)$, and $f = \prod_{i=1}^k m_{x_i,K} \in K[T]$. Then let M be a splitting field of f over L. Then, since the x_i are roots of f, M is also a splitting field for f over K. So M/K is normal.

Let M' be such that $L \subseteq M' \subseteq M$ and M'/K be normal. Then as $x_i \in M'$, the minimal polynomial $m_{x_i,K}$ splits in M'. So $M' = M$.

Any normal extension M/K must contain a splitting field for f, and by the minimality condition, M must be a splitting field. By uniqueness of splitting fields, any two such extensions are L-isomorphic as required. □

3.4. Separable polynomials

Recall that over $\mathbb{C}$, a root x of a polynomial is said to be a multiple zero when its derivative vanishes at x. Over arbitrary fields, the same is true, but the analytic concept of derivative must be replaced with an algebraic process.

Definition. The *formal derivative* of a polynomial $f(T) = \sum_{i=0}^d a_i X^i$ is

$$f'(T) = \sum_{i=1}^d i a_i X^{i-1}$$

Remark. One can check from the definition that the familiar rules $(f+g)' = f'+g'$, $(fg)' = f'g + fg'$, and $(f^n)' = nf'f^{n-1}$ hold.

Example. Consider a field K of characteristic $p > 0$, and let $f = T^p + a_0$. Then $f' = 0$, so a non-constant polynomial can have a zero derivative.

Proposition. Let $f \in K[T]$, L/K be a field extension, and $x \in L$ a root of f. Then x is a simple root if and only if $f'(x) \neq 0$.

Proof. We can write $f = (T-x)g \in L[T]$. Then $f' = g + (T-x)g'$, so $f'(x) = g(x)$. In particular, $f'(x) \neq 0$ if and only if $(T-x)$ does not divide g, which is the criterion that x is a simple root of f. □

Definition. A polynomial $f \in K[T]$ is *separable* if it splits into distinct linear factors in a splitting field. Equivalently, it has $\deg f$ distinct roots.

Corollary. f is separable if and only if the greatest common divisor of f and f' is 1.

For convenience, we will take $\gcd(f,g)$ to be the unique monic polynomial h such that $(h) = (f,g)$. Then since $K[T]$ is a Euclidean domain, we can compute a representation $h = af+bg$ for polynomials a, b. Note that $\gcd(f,g)$ is invariant under a field extension, because Euclid's algorithm does not depend on the ambient field structure.

Proof. We can replace K by a splitting field of f, so we can factorise f into a product of linear factors in K. The two are separable if f, f' have no common root, which is true if and only if $\gcd(f, f') = 1$. □

Example. Let K have characteristic $p > 0$, and let $f = T^p - b$ for $b \in K$. Then $f' = 0$, so $\gcd(f, f') = f \neq 1$. Hence f is inseparable. Let L be an extension of K containing a pth root $a \in L$ of b, so $a^p = b$. Then $f = (T - a)^p = T^p + (-a)^p = T^p - b$. In particular, f has only one root in a splitting field.

If b is not a pth power in K, then f is irreducible. This is seen on the example sheets.

Theorem. Let $f \in K[T]$ be an irreducible polynomial. Then f is separable if and only if $f' \neq 0$.

In addition, if K has characteristic zero, every irreducible polynomial $f \in K[T]$ is separable. If K has positive characteristic $p > 0$, an irreducible polynomial $f \in K[T]$ is inseparable if and only if $f(T) = g(T^p)$ for some $g \in K[T]$.

Proof. Without loss of generality, we can assume f is monic. Then, since f is irreducible, the greatest common divisor $\gcd(f, f')$ is either f or 1. If $\gcd(f, f') = f$, then $f' = 0$ by considering the degree.

For a polynomial f, we can write $f = \sum_{i=0}^{d} a_i T^i$ and $f' = \sum_{i=1}^{d} i a_i T^{i-1}$, so $f' = 0$ if and only if $i a_i = 0$ for all $1 \leq i \leq d$. In particular, if K has characteristic zero, this is true if and only if $a_i = 0$ for all $1 \leq i \leq d$, so $f = a_0$ is a constant so not irreducible. If K has characteristic $p > 0$, the requirement is that $a_i = 0$ for all i not divisible by p, or equivalently, $f(T) = g(T^p)$. □

3.5. Separable extensions

Definition. Let L/K be a field extension. We say $x \in L$ is *separable* over K if x is algebraic and its minimal polynomial f is separable over K. L is *separable* over K if all elements x are separable over K.

Theorem. Let x be algebraic over K, and L/K be an extension in which the minimal polynomial $m_{x,K}$ splits. Then x is separable over K if and only if there are exactly $\deg x$ K-homomorphisms from $K(x)$ to L.

Proof. The number of K-homomorphisms from $K(x)$ to L is the number of roots of $m_{x,K}$ in L. This is equal to the degree of x if and only if x is separable. □

Let $\mathrm{Hom}_K(L, M)$ be the set of K-homomorphisms from L to M. Note that not all K-linear maps from L to M are K-homomorphisms.

Theorem (counting embeddings)**.** Let $L = K(x_1, \dots, x_k)$ be a finite extension of K, so the x_i are algebraic. Let M/K be any field extension. Then $|\mathrm{Hom}_K(L, M)| \leq [L : K]$, with equality if and only if

(i) for all i, the minimal polynomial $m_{x_i,K}$ splits into linear factors in M; and

(ii) all the x_i are separable over K.

Remark. The conditions (i) and (ii) are equivalent to the statement that $m_{x_i,K}$ split into distinct linear factors over M. There is a variant of this theorem: let $\sigma : K \to M$ be a field homomorphism, then $|\mathrm{Hom}_\sigma(L, M)| \leq |L : K|$, and equality holds if and only if the $\sigma m_{x_i,K}$ split into distinct linear factors over M.

Proof. We prove this by induction on k. The case $k = 0$ is trivial. Let $K_1 = K(x_1)$ and write $d = \deg_K x_1 = [K_1 : K]$. Then the number of K-homomorphisms from K_1 to M, denoted $e = |\mathrm{Hom}_K(K_1, M)|$, is the number of roots of $m_{x_1,K}$ in M. Let $\sigma : K_1 \to M$ be a K-homomorphism. By the inductive hypothesis, there exist at most $[L : K_1]$ extensions of σ to a K-homomorphism $L \to M$. Hence the number of K-homomorphisms from L to M is at most $e[L : K_1] \leq d[L : K_1] = [L : K]$.

If equality holds, then $e = d$, and so $m_{x_1,K}$ splits into d distinct linear factors in M, so (i) and (ii) hold for x_1. Replacing x_1 with an arbitrary x_i, one implication follows. Conversely, suppose conditions (i) and (ii) hold. Then, by the previous theorem, there are d distinct homomorphisms from K_1 to M. Conditions (i) and (ii) still hold over K_1, then by induction, each $\sigma : K_1 \to M$ has $[L : K_1]$ extensions to a homomorphism $L \to M$. Hence $|\mathrm{Hom}_K(L, M)| = [L : K]$ as required. □

Theorem (separably generated implies separable)**.** Let $L = K(x_1, \dots, x_k)$ be a finite extension of K. Then L/K is a separable extension if and only if each x_i is separable over K.

Proof. If L/K is separable, the x_i are separable by definition. Suppose the x_i are separable. Let M be a normal closure of L/K, so the splitting field of the product of the $m_{x_i,K}$ over L. By the counting embeddings theorem, conditions (i) and (ii) are satisfied so $|\mathrm{Hom}_K(L, M)| = [L : K]$. But if $x \in L$, $L = K(x, x_1, \dots, x_k)$, so x is separable. □

Corollary. Let $x, y \in L$, and L/K a field extension. If x, y are separable over K, so are $x + y, xy, x^{-1}$ for $x \neq 0$.

Proof. Consider the fields $K(x, y)$ and $K(x)$. These are separable extensions of K. In particular, $\{x \in L \mid x \text{ separable over } K\}$ is a subfield of L. □

Theorem (primitive element theorem for separable extensions)**.** Let K be an infinite field and $L = K(x_1, \dots, x_k)$ be a finite separable extension. Then there exists $x \in L$ such that $L = K(x)$. In particular, x is separable over K.

Proof. It suffices to consider the case when $k = 2$, because if we can turn $K(x, y)$ into $K(z)$ for $z \in K(x, y)$, we can perform this inductively. Let $L = K(x, y)$ with x, y separable over K. Let $n = [L : K]$, and let M be a normal closure for L/K. Then there exist n distinct K-homomorphisms $\sigma_i : L \to M$. Let $a \in K$, and consider $z = x + ay$. We will choose a such that $L = K(z)$.

Since $L = K(x, y)$, we have $\sigma_i(x) = \sigma_j(x)$ and $\sigma_i(y) = \sigma_j(y)$ implies $i = j$. Consider $\sigma_i(z) = \sigma_i(x) + a\sigma_i(y)$. If $\sigma_i(z) = \sigma_j(z)$, we must have $\big(\sigma_i(x) - \sigma_j(x)\big) + a\big(\sigma_i(y) - \sigma_j(y)\big) = 0$. If $i \neq j$, at least one of the parenthesised terms is nonzero. Therefore there is at most one $a \in K$ such that $\sigma_i(z) = \sigma_j(z)$. Since K is infinite, there exists $a \in K$ such that all of the $\sigma_i(z)$ are distinct. But then $\deg_K z = n$, so $L = K(z)$. □

Theorem. Let L/K be an extension of finite fields. Then $L = K(x)$ for some $x \in L$.

Proof. The multiplicative group $L^\times$ is cyclic. Let x be a generator of this group. Then $L = K(x)$, since every nonzero element is a power of x. □

4. Galois theory

4.1. Field automorphisms

Definition. A bijective homomorphism from a field to itself is called an *automorphism*. The set of automorphisms of a field L forms a group $\mathrm{Aut}(L)$ under composition: $(\sigma\tau)(x) = \sigma(\tau(x))$. This is called the *automorphism group of* L. Let $S \subseteq \mathrm{Aut}(L)$. Then, we define

$$L^S = \{x \in L \mid \forall \sigma \in S, \sigma(x) = x\}$$

This is a subfield of L, known as the *fixed field of* S, since each σ is a homomorphism.

Example. Let $L = \mathbb{C}$ and σ be the complex conjugation automorphism. Then the fixed field of $\{\sigma\}$ is $\mathbb{C}^{\{\sigma\}} = \mathbb{R}$.

Definition. Let L/K be a field extension. We define $\mathrm{Aut}(L/K)$ to be the set of K-automorphisms of L, so $\mathrm{Aut}(L/K) = \{\sigma \in \mathrm{Aut}(L) \mid \forall x \in K, \sigma(x) = x\}$. Equivalently, $\sigma \in \mathrm{Aut}(L)$ is an element of $\mathrm{Aut}(L/K)$ if $K \subseteq L^{\{\sigma\}}$. $\mathrm{Aut}(L/K)$ is a subgroup of $\mathrm{Aut}(L)$.

Theorem. Let L/K be a finite extension. Then $|\mathrm{Aut}(L/K)| \leq [L : K]$.

Proof. Let $M = L$, then $\mathrm{Hom}_K(L, M) = \mathrm{Aut}(L/K)$, which has at most $[L : K]$ elements. □

Proposition. If $K = \mathbb{Q}$ or $K = \mathbb{F}_q$, $\mathrm{Aut}(K) = \{1\}$.

Proof. $\sigma(1_K) = 1_K$ hence $\sigma(n_K) = n_K$. □

In particular, $\mathrm{Aut}(L) = \mathrm{Aut}(L/K)$ where K is the prime subfield of L.

4.2. Galois extensions

We need to define a notion of when an extension L/K has 'many symmetries'.

Definition. An extension L/K is a *Galois extension* if it is algebraic, and $L^{\mathrm{Aut}(L/K)} = K$.

Remark. If $x \in L \setminus K$, there is a K-automorphism $\sigma : L \to L$ such that $x \neq \sigma(x)$.

Example. $\mathbb{C}/\mathbb{R}$ is a Galois extension, since the fixed field of complex conjugation is $\mathbb{R}$. Similarly, $\mathbb{Q}(i)/\mathbb{Q}$ is a Galois extension.

Example. Let $K/\mathbb{F}_p$ be a finite extension, so K is a finite field. The Frobenius automorphism of K, given by $\varphi_p(x) = x^p$, has fixed field

$$K^{\{\varphi_p\}} = \{x \in K \mid x \text{ a root of } T^p - T\}$$

But since this has at most p roots, and each element of $\mathbb{F}_p$ is a root, the fixed field is exactly $\mathbb{F}_p$. So $K^{\mathrm{Aut}(K/\mathbb{F}_p)} = \mathbb{F}_p$, so this is a Galois extension.

Definition. Let L/K be a Galois extension. We write $\mathrm{Gal}(L/K)$ for the automorphism group $\mathrm{Aut}(L/K)$, called the *Galois group of L/K*.

Theorem (classification of finite Galois extensions)**.** Let L/K be a finite extension, and let $G = \mathrm{Aut}(L/K)$, then the following are equivalent.

(i) L/K is a Galois extension, so $K = L^G$.

(ii) L/K is normal and separable.

(iii) L is a splitting field of a separable polynomial in K.

(iv) $|\mathrm{Aut}(L/K)| = [L : K]$.

If this holds, the minimal polynomial of any $x \in L$ over K is $m_{x,K} = \prod_{i=1}^{r}(T - x_i)$, where $\{x_1, \dots, x_r\}$ is the orbit of G on x.

Proof. (i) implies (ii) and the minimal polynomial result. Let $x \in L$, and $\{x_1, \dots, x_r\}$ be the orbit of G on x. Let $f = \prod_{i=1}^{r}(T - x_i)$. Then $f(x) = 0$. Since G permutes the x_i, the coefficients of f are fixed by G. By assumption, the coefficients of f lie in K, so the minimal polynomial of x must divide f. Since $m_{x,K}(\sigma(x)) = \sigma(m_{x,K}(x)) = 0$, so every x_i is a root of the minimal polynomial of $m_{x,K}$. So f is exactly the minimal polynomial as required. $m_{x,K}$ is a separable polynomial and splits in L. So L/K is normal and separable.

(ii) implies (iii). Since splitting fields are normal extensions, L is a splitting field for some polynomial $f \in K[T]$. Write $f = \prod_{i=1}^{r} q_i^{e_i}$ where the q_i are distinct irreducible polynomials, and $e_i \geq 1$. Since L and K are separable, the q_i are separable as they are irreducible, so $g = \prod_{i=1}^{r} q_i$ is separable and L is also a splitting field for g.

(iii) implies (iv). Let $L = K(x_1, \dots, x_k)$ be the splitting field of a separable polynomial $f \in K[T]$ with roots x_i. By the theorem on counting embeddings with $M = L$, since $m_{x_i,K} \mid f$, conditions (i) and (ii) in the theorem are satisfied, and we find $|\mathrm{Aut}(L/K)| = |\mathrm{Hom}_K(L, M)| = [L : K]$.

(iv) implies (i). Suppose $|\mathrm{Aut}(L/K)| = |G| = [L : K]$. Note that $G \subseteq \mathrm{Aut}(L/L^G) \subseteq \mathrm{Aut}(L/K)$, so these inclusions are both equalities. So $G = \mathrm{Aut}(L/L^G)$, so $[L : K] = |G| \leq [L : L^G]$. But since $L^G \supseteq K$, we must have equality by the tower law. □

Corollary. Let L/K be a finite Galois extension. Then $L = K(x)$ for some $x \in L$ which is separable over K, and has degree $[L : K]$.

Proof. By (ii) above, L/K is separable. Then the primitive element theorem implies that $L = K(x)$ for some x. □

4.3. Galois correspondence

Theorem (Galois correspondence: part (a))**.** Let L/K be a finite Galois extension with $G = \mathrm{Gal}(L/K)$. Suppose F is another field, and $K \subseteq F \subseteq L$. Then L/F is also a Galois extension where $\mathrm{Gal}(L/F) \leq \mathrm{Gal}(L/K)$. The map $F \mapsto \mathrm{Gal}(L/F)$ is a bijection between the set of intermediate fields F and the set of subgroups of $H \leq \mathrm{Gal}(L/K)$. The inverse of this map is $H \mapsto L^H$. This bijection reverses inclusions, and if $F = L^H$, we have $[F : K] = (G : H)$.

Proof. Let $x \in L$. Then $m_{x,F} \mid m_{x,K}$ in $F[T]$. As $m_{x,K}$ splits into distinct linear factors in L so does $m_{x,F}$. So L/F is normal and separable, and hence a Galois extension as required. By definition, $\mathrm{Gal}(L/F) \leq \mathrm{Gal}(L/K)$.

To check the map $F \mapsto \mathrm{Gal}(L/F)$ is a bijection with the given inverse, we first consider a field F, and its image $L^{\mathrm{Gal}(L/F)}$ under both maps. We have $L^{\mathrm{Gal}(L/F)} = F$, since L/F is Galois as required. Conversely, suppose $H \leq \mathrm{Gal}(L/F)$, and consider its image $\mathrm{Gal}(L/L^H)$. To show $\mathrm{Gal}(L/L^H) = H$, it suffices to show that $[L : L^H] \leq |H|$, because certainly $H \leq \mathrm{Gal}(L/L^H)$ and $|\mathrm{Gal}(L/L^H)| \leq [L : L^H]$. By the previous corollary, $L = L^H(x)$ for some x, and $f = \prod_{\sigma \in H}(T - \sigma(x)) \in L^H[T]$ is a polynomial with x as a root. In particular, $[L : L^H] = \deg_{L^H}(x) \leq \deg f = |H|$. So we have a bijection as claimed.

Suppose $F \subseteq F'$ are fields between K and L. Then $\mathrm{Gal}(L/F') \subseteq \mathrm{Gal}(L/F)$, so the bijection reverses inclusions. Finally, if $F = L^H$, we have $[F : K] = \frac{[L:K]}{[L:F]} = \frac{|\mathrm{Gal}(L/K)|}{|\mathrm{Gal}(L/F)|} = \frac{|G|}{|H|} = (G : H)$. □

Theorem (Galois correspondence: part (b))**.** Let $H \leq G$ be a subgroup of a Galois group $G = \mathrm{Gal}(L/K)$. Then $\sigma H \sigma^{-1}$ corresponds to the field σL^H.

Proof. Under the Galois correspondence, $\sigma H \sigma^{-1}$ corresponds to its fixed field

$$L^{\sigma H \sigma^{-1}} = \left\{x \in L \mid \sigma\tau\sigma^{-1}(x) = x \text{ for all } \tau \in H\right\}$$

Note that $\sigma\tau\sigma^{-1}(x) = x$ if and only if $\tau\sigma^{-1}(x) = \sigma^{-1}(x)$, so $\tau(y) = y$ for $x = \sigma(y)$. Hence $x \in L^{\sigma H \sigma^{-1}}$ if and only if there exists $y \in L^H$, $x = \sigma(y)$. Therefore $L^{\sigma H \sigma^{-1}} = \sigma L^H$ as required. □

Theorem (Galois correspondence: part (c))**.** Let $H \leq G = \mathrm{Gal}(L/K)$. Then the following are equivalent.

(i) L^H/K is Galois;

(ii) L^H/K is normal;

(iii) for all $\sigma \in G$, $\sigma L^H = L^H$;

(iv) H is a normal subgroup of G.

If so, $\mathrm{Gal}(L^H/K) = {}^{\mathrm{Gal}(L/K)}\!/\!_{H} = {}^{G}\!/\!_{H}$.

Proof. (i) and (ii) are equivalent. L/K is separable since it is Galois. So L^H/K is also separable.

(iii) and (iv) are equivalent. Let $F = L^H$, and let $x \in F$. Then the set of roots of $m_{x,K}$ is the orbit of x under G, so the minimal polynomial splits in F if and only if for all $\sigma \in G$, $\sigma(x) \in F$.

As this holds for all $x \in F$, F is normal if and only if $\sigma F \subseteq F$. Since $[\sigma F : K] = [F : K]$, as F and σF are K-isomorphic, this holds if and only if $\sigma F = F$. By part (b) of the Galois correspondence, this is equivalent to the statement that $\sigma H \sigma^{-1} = H$ for all σ, so H is normal.

If any of the above hold, for all $\sigma \in G$, we have $\sigma F = F$, so we have homomorphisms $G \to \mathrm{Gal}(F/K)$ given by the restriction of $\sigma \in G$ to F. Its kernel is H. Then from the isomorphism theorem, ${}^{G}\!/_{H}$ is isomorphic to a subgroup of $\mathrm{Gal}(F/K)$. This must be an isomorphism because $[F : K] = (G : H)$. □

Example. Let $K = \mathbb{Q}$ and $L = \mathbb{Q}(\sqrt[3]{2}, \omega)$ where $\omega = e^{\frac{2\pi i}{3}}$. L is a splitting field for $T^3 - 2$ with $[L : \mathbb{Q}] = 6$. Since $T^3 - 2$ is a separable polynomial, L is the splitting field of a separable polynomial and hence Galois. Therefore $G = \mathrm{Gal}(L/\mathbb{Q})$ has order 6.

We have the subfields $F_1 = \mathbb{Q}(\omega)$, $F_2 = \mathbb{Q}(\sqrt[3]{2})$, where $[F_1 : \mathbb{Q}] = 2$ and $[F_2 : \mathbb{Q}] = 3$. In the following diagram, the arrows on the left hand side are annotated with the degrees of an extensions, and the arrows on the right hand side are labelled with the index of the relevant subgroup.

$$\begin{tikzcd}
& L & \\
F_1 \arrow[ur, "3"] & & F_2 \arrow[ul, "2"'] \\
& K \arrow[ul, "2"] \arrow[ur, "3"'] \arrow[uu, "6"] &
\end{tikzcd}
\qquad\qquad
\begin{tikzcd}
& \{1\} \arrow[dl, "3"'] \arrow[dr, "2"] \arrow[dd, "6"] & \\
H_1 \arrow[dr, "2"'] & & H_2 \arrow[dl, "3"] \\
& G &
\end{tikzcd}$$

By the classification of finite groups of order 6, G is isomorphic either to C_6 or S_3. $F_2 = \mathbb{Q}(\sqrt[3]{2})$ is not a normal extension of $\mathbb{Q}$, because $\omega\sqrt[3]{2} \notin F_2$. So H_2 is not a normal subgroup of G. Since all subgroups of abelian groups are normal, G is not abelian. So $G \cong S_3$. Hence $H_1 \cong A_3$, and H_2 is a transposition, but since all subgroups generated by transpositions are conjugate, we can set $H_2 = \langle(1\ 2)\rangle$.

The other two subgroups are conjugate to H_2, corresponding to the subfields σF_2 where $\sigma \in G$. Hence, these subfields are exactly $\mathbb{Q}(\omega\sqrt[3]{2})$ and $\mathbb{Q}(\omega^2\sqrt[3]{2})$, since the conjugates of $\sqrt[3]{2}$ are exactly the roots of the minimal polynomial. Note that since these are the only subgroups, we have found all intermediate fields between $\mathbb{Q}$ and $\mathbb{Q}(\sqrt[3]{2}, \omega)$.

There is an easier way to prove $G \cong S_3$. Consider a separable polynomial $f \in K[T]$, and its roots $x_1, \dots, x_n$ in a splitting field L. Then $G = \mathrm{Gal}(L/K)$ permutes the $\{x_i\}$, because $f(\sigma x_i) = \sigma f(x_i) = 0$. If $\sigma(x_i) = x_i$ for all i, since $L = K(x_1, \dots, x_n)$, σ must be the identity map. This gives an injective homomorphism from G into S_n. So G is isomorphic to a subgroup of S_n. In our example above, $|G| = 6$ and G is isomorphic to a subgroup of S_3, so $G \cong S_3$.

4.4. Galois groups of polynomials

Definition. Let $f \in K[T]$, and let L be a splitting field for f. There is an action of $\mathrm{Gal}(L/K)$ on the set of roots of f in L. If f has n roots, this action induces a subgroup of permutations of roots $\mathrm{Gal}(f/K) \leq S_n$, called the *Galois group of f over K*.

Remark. $\mathrm{Gal}(f/K) \simeq \mathrm{Gal}(L/K)$ as L is a splitting field for f over K. In particular, $[L : K] = |\mathrm{Gal}(L/K)| = |\mathrm{Gal}(f/K)| \mid n!$.

There exist several methods for finding the Galois group for a particular polynomial.

Proposition. $f \in K[T]$ is irreducible if and only if $\mathrm{Gal}(f/K)$ is *transitive*, so for all $i, j \in \{1, \dots, n\}$, there exists $\sigma \in \mathrm{Gal}(f/K)$ such that $\sigma(i) = j$.

Remark. A subgroup of S_n is transitive if and only if there is exactly one orbit.

Proof. Let x be a root of f in a splitting field L. Then its orbit under $G = \mathrm{Gal}(f/K)$ is exactly the set of roots of $m_{x,K}$. Since $m_{x,K}$ is an irreducible factor of f, $m_{x,K} = f$ if and only if f is irreducible. Conversely, $m_{x,K} = f$ if and only if each root of f is in the orbit of x, which is exactly the statement that G acts transitively on the roots of f. □

Remark. If $G \subseteq S_n$ is transitive, by the orbit-stabiliser theorem, $n \mid |G|$.

Recall that for a monic polynomial $f = \prod_{i=1}^{n}(T - x_i)$, the *discriminant* of f is $\mathrm{Disc}(f) = \Delta^2 \in K$, where $\Delta = \prod_{i<j}(x_i - x_j)$. The discriminant is nonzero if and only if f is separable.

Proposition. Let $\operatorname{char} K \neq 2$, and let $f \in K[T]$ be a monic polynomial with splitting field L. Let $G = \mathrm{Gal}(f/K)$. Then the fixed field of $G \cap A_n$ is $K(\Delta)$, where Δ^2 is the discriminant. In particular, $\mathrm{Gal}(f/K) \subseteq A_n$ if and only if the discriminant $\mathrm{Disc}(f)$ is a square.

Proof. Let $\pi \in S_n$. The sign of the permutation is given by

$$\prod_{i<j}(T_{\pi(i)} - T_{\pi(j)}) = \operatorname{sgn} \pi \prod_{i<j}(T_i - T_j)$$

Hence, if $\sigma \in G$, we have $\sigma(\Delta) = \operatorname{sgn} \sigma \cdot \Delta$. Because the characteristic is not 2, $-1 \neq 1$. Since $\Delta \neq 0$, this implies $\Delta \in K$ if and only if $G \subseteq A_n$, and Δ lies in the fixed field F of $G \cap A_n$. Because $[F : K] = (G : G \cap A_n) \in \{1, 2\}$, $F = K(\Delta)$ exactly. □

Example. Let $n = 3$, $f = T^3 + aT + b = \prod_{i=1}^{3}(T - x_i)$ where x_i lie in a splitting field for f. Since there is no T^2 term, $x_3 = -x_1 - x_2$. Hence, $a = x_1x_2 - (x_1 + x_2)^2$, and $b = x_1x_2(x_1 + x_2)$. Therefore,

$$\mathrm{Disc}(f) = \left[(x_1 - x_2)(2x_1 + x_2)(x_1 + 2x_2)\right]^2 = -4a^3 - 27b^2$$

In particular, $\mathrm{Gal}(f/K) \subseteq A_3$ if and only if $-4a^3 - 27b^2$ is a square in K.

For example, consider $f = T^3 - 21T - 7 \in \mathbb{Q}[T]$. This is irreducible by Eisenstein's criterion. Its discriminant is $4 \cdot 21^3 - 27 \cdot 7^2 = (27 \cdot 7)^2$, which is a square. So $\mathrm{Gal}(f/K) \subseteq A_3$. Since f is irreducible, $\mathrm{Gal}(f/K)$ is transitive, so its order is divisible by 3. So $\mathrm{Gal}(f/K)$ must be exactly A_3.

Remark. This technique can be used to calculate the Galois group of any cubic polynomial for characteristic not $2, 3$, for example.

5. Finite fields

5.1. Construction of finite fields

Every finite field has characteristic $p > 0$, and so it can be regarded as a field extension of $\mathbb{F}_p$. We will classify every finite field and study their Galois theory. Recall that, for a finite field F of characteristic p,

(i) $|F| = p^n$, where $[F : \mathbb{F}_p] = n$;

(ii) $F^\times$ is cyclic, of order $p^n - 1$;

(iii) The Frobenius automorphism $\varphi_p : F \to F$ given by $x \mapsto x^p$ is an automorphism of F.

Theorem. Let p be a prime, and $n \geq 1$. Then there is a finite field with $q = p^n$ elements. Any such field is a splitting field of the polynomial $f = T^q - T$ over $\mathbb{F}_p$. Since splitting fields are unique up to $\mathbb{F}_p$-isomorphism, any two finite fields of the same order are isomorphic.

Proof. Let F be a field with $q = p^n$ elements. Then if $x \in F^\times$, $x^{q-1} = 1$. Hence, for all $x \in F$, $x^q = x$. In particular, f has q distinct roots in F, which are all of the elements of F. So f splits into linear factors in F, and not in any proper subfield, so F is indeed a splitting field for f as required.

Now, we wish to explicitly construct such a field. Let L be a splitting field for $f = T^q - T$ over $\mathbb{F}_p$. Let $F \subseteq L$ be the fixed field of φ_p^n, the map $x \mapsto x^q$. So F is the set of roots of f in L. So $|F| = q$. Therefore, $L = F$ because F has q elements, using the above argument. □

Now that we have shown isomorphism, we simply write $\mathbb{F}_q$ for any finite field of q elements. There is no canonical finite field of a given order in general.

5.2. Galois theory of finite fields

Theorem. The extension $\mathbb{F}_{p^n}/\mathbb{F}_p$ is Galois, and the Galois group is cyclic of order n, generated by the Frobenius automorphism φ_p.

Proof. Since $\mathbb{F}_{p^n}$ is the splitting field of the separable polynomial $T^{p^n} - T$, the extension is Galois. Let $G \leq \mathrm{Gal}(\mathbb{F}_{p^n}/\mathbb{F}_p)$ be the subgroup generated by φ_p. Then $\mathbb{F}_{p^n}^G = \{x \mid x^p = x\} = \mathbb{F}_p$, so by the Galois correspondence, G must be the entire group $\mathrm{Gal}(\mathbb{F}_{p^n}/\mathbb{F}_p)$. □

Theorem. $\mathbb{F}_{p^n}$ has a unique subfield of order p^m for all $m \mid n$, and no others. If $m \mid n$, then $\mathbb{F}_{p^m} \subseteq \mathbb{F}_{p^n}$ is the fixed field of φ_p^m.

Proof. By the Galois correspondence, it suffices to check the subgroups of $\mathbb{Z}/n\mathbb{Z}$. The subgroups of $\mathbb{Z}/n\mathbb{Z}$ are $m\mathbb{Z}/n\mathbb{Z}$ for $m \mid n$. Hence, the subfields of $\mathbb{F}_{p^m}$ are the fixed fields of the subgroups $\langle \varphi_p^m \rangle$, which have degree equal to the indices $\left(\mathbb{Z}/n\mathbb{Z} : m\mathbb{Z}/n\mathbb{Z}\right) = m$. □

Remark. If $m \mid n$, $\operatorname{Gal}(\mathbb{F}_{p^n}/\mathbb{F}_{p^m}) = \langle \varphi_p^m \rangle$, which has order $\frac{n}{m}$.

Theorem. Let $f \in \mathbb{F}_p[T]$ be separable, and let $n = \deg f$. Suppose the irreducible factors of f have degrees $n_1, \dots, n_r$, so $\sum_{i=1}^r n_i = n$. Then $\operatorname{Gal}(f/\mathbb{F}_p) \subseteq S_n$ is cyclic and generated by an element of cycle type $(n_1, \dots, n_r)$. In particular, $|\operatorname{Gal}(f/\mathbb{F}_p)|$ is the least common multiple of the n_i.

Recall that $\pi \in S_n$ has cycle type $(n_1, \dots, n_r)$ if it is a product of r disjoint cycles π_i, each with length n_i.

Proof. Let L be a splitting field for f over $\mathbb{F}_p$. Consider $x_1, \dots, x_n \in L$. Then $\operatorname{Gal}(L/\mathbb{F}_p)$ is cyclic and generated by φ_p. As the irreducible factors g_i of f are the minimal polynomials of the x_i, and the roots of the minimal polynomial of x_i are precisely the orbit of φ_p on x_i, the cycle type must be as required. The order of any such permutation is the lowest common multiple of the lengths of the cycles. □

5.3. Reduction modulo a prime

Theorem. Let $f \in \mathbb{Z}[T]$ be a monic separable polynomial with $\deg f = n$, and let p be a prime. Suppose that the *reduction* $\overline{f} \in \mathbb{F}_p[T]$ of f is also separable. Then $\operatorname{Gal}(\overline{f}/\mathbb{F}_p) \leq \operatorname{Gal}(f/\mathbb{Q})$ as subgroups of S_n.

Remark. The identification of $\operatorname{Gal}(f/\mathbb{Q})$ with a subgroup of S_n depends on the choice of ordering of the roots of f. Choosing a different ordering corresponds to conjugation of $\operatorname{Gal}(f/\mathbb{Q})$ in S_n. The meaning of the statement $\operatorname{Gal}(\overline{f}/\mathbb{F}_p) \leq \operatorname{Gal}(f/\mathbb{Q})$ therefore means that $\operatorname{Gal}(\overline{f}/\mathbb{F}_p)$ is conjugate to a subgroup of $\operatorname{Gal}(f/\mathbb{Q})$ in S_n, not that it is exactly a subgroup.

The following proof is based in algebraic number theory; alternatives are available. The proof is not examinable.

Proof. Let $L = \mathbb{Q}(x_1, \dots, x_n)$ be a splitting field for f, where the x_i are the roots of f. Let $N = [L : \mathbb{Q}]$. Consider $R = \mathbb{Z}[x_1, \dots, x_n]$. Since $f(x_i) = 0$ and f is monic, every element of R is a $\mathbb{Z}$-linear combination of $x_1^{a_1}, \dots, x_n^{a_n}$ where the $a_i < n$ by using f to reduce the degrees. So R is finitely-generated as a $\mathbb{Z}$-module, or equivalently, as an abelian group. R is contained inside $L \simeq \mathbb{Q}^N$. R is torsion-free, so $R \simeq \mathbb{Z}^M$ with $M \leq N$ (in fact, $M = N$).

Then $\overline{R} = {}^{R}\!/_{pR}$ has p^M elements. Let $\overline{P}$ be a maximal ideal for $\overline{R}$, which corresponds to an ideal P of R that contains pR. Then $F = {}^{R}\!/_{P} \simeq {}^{\overline{R}}\!/_{\overline{P}}$ (by the isomorphism theorem) is a finite field with p^d elements for some d. Since R is generated by $x_1, \dots, x_n$, F is generated by $\overline{x}_1, \dots, \overline{x}_n$, where $\overline{x}_i = x_i + P \in F$. In particular, $\overline{f} = \prod_{i=1}^n (T - \overline{x}_i)$. Since $\overline{f}$ is separable, the $\overline{x}_i$ are distinct, and F is a splitting field for $\overline{f}$.

Let $G = \operatorname{Gal}(f/\mathbb{Q})$. Then G maps R to R since it permutes the x_i. Let $H \leq G$ be the stabiliser of P, so $H = \{\sigma \in G \mid \sigma P = P\}$. Since H fixes P, H acts on the quotient ${}^{R}\!/_{P} = F$, and it permutes the $\overline{x}_i$ in the same way as it permutes the x_i. In particular, there is an injective homomorphism from H into $\mathrm{Gal}(F/\mathbb{F}_p)$. It now suffices to show that this homomorphism is an isomorphism.

Let $\{P = P_1, P_2, \dots, P_r\}$ be the orbit of P under G, so $P_i = \sigma P$ for some $\sigma \in G$. These are all maximal ideals since P is, and ${}^{R}\!/\!{}_{P_i} \simeq {}^{R}\!/\!{}_{P}$ so each ${}^{R}\!/\!{}_{P_i}$ have p^d elements. The P_i are maximal, so $P_i + P_j = R$ if $i \neq j$. So by the Chinese remainder theorem for rings,

$$ {}^{R}\!/\!{}_{(P_1 \cap \dots \cap P_k)} \simeq {}^{R}\!/\!{}_{P_1} \times \dots \times {}^{R}\!/\!{}_{P_r} $$

As $p \in P_1$, $pR \subseteq P_1 \cap \dots \cap P_r$. So

$$ p^N \geq p^M = \left|{}^{R}\!/\!{}_{pR}\right| \geq \left|{}^{R}\!/\!{}_{(P_1 \cap \dots \cap P_r)}\right| = \prod_{i=1}^{r} \left|{}^{R}\!/\!{}_{P_i}\right| = p^{rd} \implies N \geq rd $$

Now, by the orbit-stabiliser theorem, $r = (G : H) = \frac{N}{|H|}$. Since H injects into $\mathrm{Gal}(F/\mathbb{F}_p)$, we have $|H| \leq d$ with equality if and only if the injection is an isomorphism. So $N \leq rd$, but since $N \geq rd$, we must have $N = rd$, so the injection is an isomorphism, and $H \simeq \mathrm{Gal}(\overline{f}/\mathbb{F}_p)$. □

Corollary. Let $f \in \mathbb{Z}[T]$ be monic and separable with p a prime such that $\overline{f} \in \mathbb{F}_p[T]$ is separable. Consider the factorisation into irreducibles $\overline{f} = g_1 \dots g_r \in \mathbb{F}_p[T]$, where $\deg g_i = n_i$. Then $\mathrm{Gal}(f/\mathbb{Q})$ contains an element of cycle type $(n_1, \dots, n_r)$.

Proof. Combine the previous two theorems. □

Example. Let $f = T^4 - 3T + 1$. Consider $p = 2$. In $\mathbb{F}_2$, $f = T^4 + T + 1$. This does not have a root, and not divisible by $T^2 + T + 1$ which is the only irreducible quadratic, so it is irreducible.

Now, consider $p = 5$. In $\mathbb{F}_5$, $f = (T + 1)(T^3 - T^2 + T + 1)$, which is a factorisation into irreducibles.

By the above corollary, $\mathrm{Gal}(f/\mathbb{Q})$ has a 4-cycle and a 3-cycle. In particular, $12 \mid |\mathrm{Gal}(f/\mathbb{Q})|$, so the group is either all of S_4 or it is A_4, as this is the unique index 2 subgroup of S_4. But 4-cycles are odd, so do not lie in A_4. So $\mathrm{Gal}(f/\mathbb{Q}) = S_4$.

Note that if $\overline{f}$ is separable, $\mathrm{Disc}(\overline{f}) \neq 0$, so $p \nmid \mathrm{Disc}(\overline{f})$ so f is separable. If f is separable, then $\overline{f}$ is separable for all primes but the finite set of primes dividing $\mathrm{Disc}(f)$.

Remark. If $\mathrm{Gal}(f/\mathbb{Q})$ contains an element of cycle type $(n_1, \dots, n_r)$, it can in fact be shown that there exist infinitely many primes p such that $\overline{f}$ factors into irreducibles of degrees $n_1, \dots, n_r$ in $\mathbb{F}_p$. This is known as the Chebotarev density theorem, which is a generalisation of Dirichlet's theorem on primes in arithmetic progression. However, the proof is far outside the scope of this course.

6. Cyclotomic and Kummer extensions

6.1. Primitive roots of unity

Lemma. Let C be a cyclic group of order $n > 1$. Let $a \in \mathbb{Z}$ be coprime with n, also written $(a, n) = 1$. Then the map $[a] : C \to C$ given by $[a](g) = g^a$ is an automorphism of C, and the map $\left(\mathbb{Z}/n\mathbb{Z}\right)^\times \to \mathrm{Aut}(C)$ defined by $a \mapsto [a]$ is an isomorphism.

Proof. $[a]$ is clearly a homomorphism, and since a is coprime to n, it is an automorphism since there exists b such that ab is congruent to 1 modulo n. Hence, there is an injection $\left(\mathbb{Z}/n\mathbb{Z}\right)^\times \to \mathrm{Aut}(C)$ given by $a \mapsto [a]$, and it is a homomorphism. If $\varphi \in \mathrm{Aut}(C)$ and g is a generator for C, $\varphi(g) = g^a$ for some $a \in \left(\mathbb{Z}/n\mathbb{Z}\right)^\times$. So $\varphi = [a]$, and in particular, the map is an isomorphism. □

Let K be a field and $n \geq 1$. We define $\mu_n(K) = \{x \in K \mid x^n = 1\}$ for the group (under multiplication) of nth roots of unity in K. This is a finite subgroup of $K^\times$, hence it is cyclic. The order of any element divides n, so it has order dividing n.

We say that $\zeta \in \mu_n(K)$ is a *primitive* nth root of unity if its order is exactly n. Such a ζ exists if and only if $\mu_n(K)$ has n elements, and then ζ is a generator for the group. In particular, $f = T^n - 1$ has n distinct roots, ζ^i for $i \in \{0, \dots, n-1\}$, and hence it is separable. In general, $f = T^n - 1$ is separable if and only if f is coprime with $f' = nT^{n-1}$, which holds if and only if $n \neq 0$. In this section, we assume that the characteristic of K is zero or is a positive number p that does not divide n, so f is separable.

Let L/K be a splitting field for $T^n - 1$. This is Galois since f is separable, so we can define $G = \mathrm{Gal}(L/K)$. Then $|\mu_n(L)| = n$, and so there exists a primitive nth root of unity $\zeta = \zeta_n \in L$. Such an L is called a *cyclotomic extension*.

Proposition. Let $L = K(\zeta)$. There exists an injective homomorphism $\chi = \chi_n : \mathrm{Gal}(L/K) \to \left(\mathbb{Z}/n\mathbb{Z}\right)^\times$ such that $\chi(\sigma) = a$ implies $\sigma(\zeta) = \zeta^a$. In particular, G is abelian. χ is an isomorphism if and only if G acts transitively on the set of primitive roots of unity in L.

The homomorphism χ is called the *cyclotomic character*.

Proof. $\mu_n(L)$ is cyclic and generated by ζ, so the roots of $T^n - 1$ are the powers of ζ, so $L = K(1, \zeta, \zeta^2, \dots, \zeta^{n-1}) = K(\zeta)$. Consider the action of G on L. This action permutes $\mu_n(L)$, and if $\zeta, \zeta' \in \mu_n(L)$ and $\sigma \in G$, then $\sigma(\zeta\zeta') = \sigma(\zeta)\sigma(\zeta')$, so σ acts as an automorphism of $\mu_n(L)$. $\sigma(\zeta) = \zeta$ if and only if σ is the identity because $L = K(\zeta)$. This gives an injective homomorphism $G \hookrightarrow \mathrm{Aut}\left(\mu_n(L)\right) \simeq \left(\mathbb{Z}/n\mathbb{Z}\right)^\times$.

ζ_n^a is primitive if and only if a is coprime to n. Therefore the set of primitive nth roots of unity is $\left\{\zeta^a \,\middle|\, a \in \left(\mathbb{Z}/n\mathbb{Z}\right)^\times\right\}$, which by the previous part, is the orbit of ζ under G. The map is surjective if and only if there is one orbit, so the result follows. □

6.2. Cyclotomic polynomials

Definition. Let K have characteristic zero or a prime p that does not divide n. The *nth cyclotomic polynomial* is

$$\Phi_n(t) = \prod_{a \in \left(\mathbb{Z}/n\mathbb{Z}\right)^\times} (T - \zeta_n^a)$$

in a splitting field L of $T^n - 1$.

This is the polynomial where the roots are the primitive nth roots of unity. As G permutes the primitive nth roots of unity in L, Φ_n has coefficients in $L^G = K$. The last part of the above proposition shows that χ is surjective if and only if $\Phi_n \in K[T]$ is irreducible.

$x \in L$ satisfies $x^n - 1 = 0$ if and only if x is a primitive dth root of unity for some unique $d \mid n$. Hence $T^n - 1 = \prod_{d \mid n} \Phi_d$, since the sets of roots are equal. In particular, we could have inductively defined the cyclotomic polynomials by $\Phi_n = \frac{T^n - 1}{\prod_{d \mid n, d \neq n} \Phi_d}$. This shows that the Φ_n do not depend on the choice of field K, since Φ_n is the image in $K[T]$ of a polynomial in $\mathbb{Z}[T]$.

For example, $\Phi_p = \frac{T^p - 1}{T - 1} = T^{p-1} + T^{p-2} + \dots + \dots + T + 1$. We also have $\Phi_1 = T - 1$ and $\Phi_{p^n}(T) = \frac{T^{p^n} - 1}{T^{p^{n-1}} - 1} = \Phi_p(T^{p^{n-1}})$. We have $\deg \Phi_n = \left| \left(\mathbb{Z}/n\mathbb{Z}\right)^\times \right| = \varphi(n)$ where φ is the Euler totient function.

Theorem (rationals). Let $K = \mathbb{Q}$. Then χ_n is an isomorphism for all $n > 1$. In particular, $[\mathbb{Q}(\zeta_n) : \mathbb{Q}] = \varphi(n)$, and Φ_n is irreducible over $\mathbb{Q}$.

Proof. The statements in the theorem are all equivalent by the previous results, so it suffices to prove that Φ_n is irreducible over $\mathbb{Q}$. If n is prime, we have already proven its irreducibility by Eisenstein's criterion and Gauss' lemma. We can easily extend this to the case where n is a prime power.

Note that χ_n is an isomorphism if for all primes $p \nmid n$, the residue class of $p \in \left(\mathbb{Z}/n\mathbb{Z}\right)^\times$ is in the image of χ, by factorising a as a product of primes if a is coprime to n. Let f be the minimal polynomial of ζ over $\mathbb{Q}$, and let g be the minimal polynomial of ζ^p over $\mathbb{Q}$. If $f = g$, then ζ^p lies in the orbit of $\mathrm{Gal}(L/K)$ on ζ, so p lies in the image of χ as required. Otherwise, f and g are coprime, and they divide $T^n - 1$ so $fg \mid T^n - 1$. As ζ is a root of $g(T^p)$, we have $f \mid g(T^p)$. Reducing modulo p, $\overline{f} \in \mathbb{F}_p[T]$ divides $\overline{g(T^p)} \in \mathbb{F}_p[T]$. But since we are working over $\mathbb{F}_p$, $\overline{g(T^p)} = \overline{g}(T)^p$. Now, $\overline{f}$ and $\overline{g}$ divide $T^n - 1$ in $\mathbb{F}_p[T]$, which is separable because $p \nmid n$. So $\overline{f} \mid \overline{g}^p$, so $\overline{f} \mid \overline{g}$. But then $\overline{f}^2 \mid \overline{fg} \mid T^n - 1$, contradicting separability of $T^n - 1$. □

Therefore, the minimal polynomial of $e^{\frac{2\pi i}{n}}$ over $\mathbb{Q}$ is Φ_n.

Theorem (finite fields). Let $K = \mathbb{F}_p$, and let n be coprime to p. Let L be a splitting field for $T^n - 1$. Then χ_n is an isomorphism from $\mathrm{Gal}(L/K)$ to $\langle p \rangle \leq \left(\mathbb{Z}/n\mathbb{Z}\right)^\times$, the subgroup generated by the residue class of p, and $\chi_n(\varphi_p) = p \bmod n$ where φ_p is the Frobenius endomorphism $x \mapsto x^p$, which is a generator of $\mathrm{Gal}(L/K)$. Further, $[L : K] = r$, where r is the order of p modulo n. Finally, φ_p has cycle type $(r, \dots, r)$ acting as a permutation of the roots of the cyclotomic polynomial Φ_n, which are the primitive nth roots of unity.

Proof. Since $\varphi_p(\zeta) = \zeta^p$ and $L = K(\zeta)$, by definition of χ_n, we have $\chi_n(\varphi_p) = p$, or more precisely, $p \bmod n$. In particular, $\chi_n(G) = \langle p \rangle$, and as this is a Galois extension, $[L : K] = |G| = |\langle g \rangle| = r$. For the last part, notice that if a and n are coprime, $\varphi_p^k(\zeta^a) = \zeta^a$ holds if and only if $\varphi_p^k(\zeta) = \zeta$, or equivalently, $r \mid k$. So the orbits of φ_p on the set $\{\zeta_n^a \mid (a, n) = 1\}$, which is the set of roots of Φ_n, all have length r. □

Remark. This almost gives another proof of the irreducibility of the cyclotomic polynomials Φ_n over $\mathbb{Q}$. By reduction modulo p, $\mathrm{Gal}(\Phi_n/\mathbb{Q})$ contains $\mathrm{Gal}(\Phi_n/\mathbb{F}_p)$ as a subgroup, up to conjugacy by elements of $S_{\varphi(n)}$. It is not difficult to show that in fact $\chi_n(\mathrm{Gal}(\Phi_n/\mathbb{Q})) \supseteq \chi_n(\mathrm{Gal}(\Phi_n/\mathbb{F}_p)) = \langle p \rangle$. As this holds for all primes p not dividing n, $\chi_n(\mathrm{Gal}(\Phi_n/\mathbb{Q})) = \left(\mathbb{Z}/n\mathbb{Z}\right)^\times$.

Remark. The last part of the above theorem implies that over $\mathbb{F}_p$, the cyclotomic polynomial Φ_n factors as a product of irreducibles of degree r. This depends only on the value of p modulo n. In general, for a polynomial with integer coefficients $f \in \mathbb{Z}[T]$, its factorisation modulo p does not follow an obvious pattern.

Answering this question is part of the Langlands programme, a large area of research in modern number theory. The case where there is such a congruence pattern turns out to be when $\mathrm{Gal}(f/\mathbb{Q})$ is abelian. This study is known as class field theory, which is studied in Part III.

6.3. Quadratic reciprocity

The following theorem is from Part II Number Theory. This theorem has several hundred proofs, and this particular one follows from the above theory on cyclotomic polynomials.

Let p be an odd prime and a an integer coprime to p. Then the *Legendre symbol* $\left(\frac{a}{p}\right)$ is defined by

$$\left(\frac{a}{p}\right) = \begin{cases} +1 & \text{if } a \text{ is a square mod } p \\ -1 & \text{otherwise} \end{cases}$$

Euler's formula for the Legendre symbol is

$$\left(\frac{a}{p}\right) \equiv a^{\frac{p-1}{2}} \mod p$$

Let q be another odd prime, and consider the case $n = q$ in the above discussion, so $L = K(\zeta_q)$ is a splitting field for $f = T^q - 1 = (T-1)\Phi_q$. On roots of f in L, the Frobenius map φ_p has cycle type $(1, r, \dots, r)$. There are $\frac{q-1}{r}$-many r-cycles. The sign of the permutation φ_p

is $(-1)^{(r-1)\frac{q-1}{r}} = (-1)^{\frac{q-1}{r}}$ since q is odd. Note that $2 \mid \frac{q-1}{r}$ holds if and only if $r \mid \frac{q-1}{2}$, or equivalently, $p^{\frac{q-1}{2}} \equiv 1 \bmod 2$. This is in the form of Euler's formula for the Legendre symbol. So the sign of φ_p is exactly $\left(\frac{p}{q}\right)$.

Since $G = \langle \varphi_p \rangle$, the sign of φ_p is $+1$ if and only if $G \subseteq A_q$ since $q = \deg f$. This holds if and only if $\mathrm{Disc}(f)$ is a square in $\mathbb{F}_p$.

Lemma. Let $f = \prod(T - x_i)$ over some field. Then $\mathrm{Disc}(f) = (-1)^{\frac{d(d-1)}{2}} \prod f'(x_i)$. where $d = \deg f$.

This lemma can be shown directly from the definition of the discriminant. We use the above lemma with $f = T^q - 1 = \prod_{a=0}^{q-1}(T - \zeta_q^n)$ and $f' = qT^{q-1}$ to find

$$\mathrm{Disc}(f) = (-1)^{\frac{q(q-1)}{2}} \prod_{a=0}^{q-1} q\zeta_q^{a(q-1)} = (-1)^{\frac{q-1}{2}} q^q \zeta_q^{(q-1)\frac{q(q-1)}{2}} = (-1)^{\frac{q-1}{2}} q^q$$

since q is odd. Hence, by the fact that $\left(\frac{-1}{p}\right) = (-1)^{\frac{p-1}{2}}$,

$$\left(\frac{p}{q}\right) = \left(\frac{\mathrm{Disc}(f)}{p}\right) = \left(\frac{(-1)^{\frac{q-1}{2}} q}{p}\right) = \left(\frac{q}{p}\right)(-1)^{\frac{(p-1)(q-1)}{4}}$$

which is the quadratic reciprocity law.

6.4. Construction of regular polygons

Lemma. If m is a positive integer such that $2^m + 1$ is prime, then m is a power of two.

Proof. If q is odd, $2^{qr}+1 = (2^r+1)(2^{qr-r}-2^{qr-2r}+\dots+1)$, which is a nontrivial factorisation. □

Ruler and compass construction of a regular n-gon for $n \geq 3$ is equivalent to constructing the real number $\cos\left(\frac{2\pi}{n}\right)$.

Theorem (Gauss). A regular n-gon is contructible if and only if n is a power of two multiplied by a product of distinct primes of the form $2^{2^k} + 1$.

Remark. Let $F_k = 2^{2^k} + 1$ be the kth *Fermat number*. $F_1 = 5$, $F_2 = 17$, $F_3 = 257$, and $F_4 = 65537$ are all prime. Fermat conjectured that all F_k are prime. This is false; Euler proved that $F_5 = 641 \cdot 6700417$. Many Fermat numbers are known to be composite, and no more have been found to be prime.

Proof. Recall that a real number $x \in \mathbb{R}$ is constructible if and only if there is a sequence of fields $\mathbb{Q} = K_0 \subset K_1 \subset \dots \subset K_n$ such that $x \in K_n$ and $[K_{i+1} : K_i] = 2$. In particular, if x is constructible, $[\mathbb{Q}(x) : \mathbb{Q}] = \deg_{\mathbb{Q}}(x)$ is a power of two. Note that

$$x = \cos\left(\frac{2\pi}{n}\right) = \frac{1}{2}(\zeta_n + \zeta_n^{-1}) \implies \zeta_n^2 - 2x\zeta_n + 1 = 0$$

Since $x \in \mathbb{R}$ and $\zeta_n \notin \mathbb{R}$ (for $n \geq 3$), $[\mathbb{Q}(\zeta_n) : \mathbb{Q}(x)] = 2$. If x is constructible, then $[\mathbb{Q}(\zeta_n) : \mathbb{Q}]$ is a power of two. But $[\mathbb{Q}(\zeta_n) : \mathbb{Q}] = \varphi(n)$.

Let $n = \prod_{i=1}^r p_i^{e_i}$ be the prime factorisation of n. Then $[\mathbb{Q}(\zeta_n) : \mathbb{Q}] = \prod_{i=1}^r p_i^{e_i - 1}(p - 1)$. This is a power of two if and only if for all odd p_i, we have $e_i = 1$ and $p_i - 1$ is a power of two. By the previous lemma, $\varphi(n)$ is a power of two if and only if n is of the required form.

Now suppose n is of the required form, so $\varphi(n) = 2^m$. $\mathbb{Q}(\zeta_n)/\mathbb{Q}$ is Galois, with Galois group $G \simeq \left(\mathbb{Z}/n\mathbb{Z}\right)^\times$, which has 2^m elements. There exist subgroups $G = H_0 \supset H_1 \supset \dots \supset H_m = 1$ such that $[H_i : H_{i+1}] = 2$. Indeed, as $2 \mid 2^m$, by Cauchy's theorem there exists an element $\sigma \in G$ of order 2, assuming G is not the trivial group. Take $H_{m-1} = \langle \sigma \rangle$, and then consider $G/\langle \sigma \rangle$, which contains a subgroup of order 2 by the same argument; we can proceed inductively. Then the tower of fixed fields $K_i = \mathbb{Q}(\zeta_n)^{H_i}$ is a tower of quadratic extensions by the Galois correspondence. □

6.5. Kummer extensions

Theorem (linear independence of field embeddings). Let K, L be fields. Let $\sigma_1, \dots, \sigma_n : K \to L$ be distinct field homomorphisms. Let $y_1, \dots, y_n \in L$ be such that for all $x \in K^\times$, $y_1\sigma_1(x) + \dots + y_n\sigma_n(x) = 0$. Then all $y_i = 0$. In other words, $\sigma_1, \dots, \sigma_n$ are L-linearly independent elements of the set of functions $K \to L$, considered as an L-vector space.

This is a special case, using $G = K^\times$, of the following theorem.

Theorem (linear independence of characters). Let G be a group and L be a field. Let $\sigma_1, \dots, \sigma_n : G \to L^\times$ be distinct group homomorphisms. Then $\sigma_1, \dots, \sigma_n$ are L-linearly independent elements.

Proof. We use induction on n. If $n = 1$, the result is clear. Suppose $n > 1$. Let $y_1, \dots, y_n \in L$ be such that for all $g \in G$, $y_1\sigma_1(g) + \dots + y_n\sigma_n(g) = 0$. Since the homomorphisms are distinct, there is an element $h \in G$ such that $\sigma_1(h) \neq \sigma_n(h)$. The σ_i are homomorphisms, so

$$y_1\sigma_1(hg) + \dots + y_n\sigma_n(hg) = y_1\sigma_1(h)\sigma_1(g) + \dots + y_n\sigma_n(h)\sigma_n(g) = 0$$

Multiplying the original expression in g by $\sigma_n(h)$ and subtracting,

$$y_1'\sigma_1(g) + \dots + y_{n-1}'\sigma_{n-1}(g) = 0; \quad y_i' = y_i(\sigma_i(h) - \sigma_n(h))$$

By induction, all $y_i' = 0$. But $\sigma_1(h) \neq \sigma_n(h)$, so $y_1 = 0$. So the original equation $y_1\sigma_1(g) + \dots + y_n\sigma_n(g) = 0$ can be simplified into $y_2\sigma_2(g) + \dots + y_n\sigma_n(g) = 0$, so again by induction, all y_i are zero. □

We now consider extensions of the form $L = K(x)$ for $x^n = a \in K$. The special case $a = 1$ gives the cyclotomic extensions. These extensions are not necessarily Galois; for example, $\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q}$ is not Galois. In this section, let $n > 1$, and $n \neq 0$ in K.

Theorem. Let K be a field that contains a primitive nth root of unity $\zeta = \zeta_n$. Let L/K be a field extension with $L = K(x)$, where $x^n = a \in K^\times$. Then L/K is a splitting field for $f = T^n - a$, and is Galois with cyclic Galois group. $[L : K]$ is the least $m \geq 1$ such that $x^m \in K$.

Proof. Note that $\boldsymbol{\mu}_n(K) = \left\{\zeta^i \mid 0 \leq i < n\right\}$ has n elements. Then f has n distinct roots $\zeta^i x$ in L. So L is a splitting field for the separable polynomial f, and in particular, L is a Galois extension.

Let $\sigma \in \mathrm{Gal}(L/K) = G$. Then $f(\sigma(x)) = 0$, so $\sigma(x) = \zeta^i x$ for some i, which is unique modulo n. This induces a map $\theta \colon G \to \boldsymbol{\mu}_n(K) \simeq \mathbb{Z}/n\mathbb{Z}$, given by $\theta(\sigma) = \frac{\sigma(x)}{x}$ which is equal to ζ^i for some i. We claim this is a homomorphism. Let $\sigma, \tau \in G$. Then since $\zeta \in K$, $\tau(\theta(\sigma)) = \theta(\sigma)$. So

$$\theta(\tau\sigma) = \frac{\tau\sigma(x)}{x} = \tau\left(\frac{\sigma(x)}{x}\right) \cdot \frac{\tau(x)}{x} = \tau(\theta(\sigma)) \cdot \theta(\tau) = \theta(\sigma)\theta(\tau)$$

It is injective, because $\theta(\sigma) = 1$ if and only if $\sigma(x) = x$, so $\sigma = \mathrm{id}$. So G is isomorphic to a subgroup of a cyclic group. Hence it is cyclic.

If $m \geq 1$, since L/K is Galois, $x^m \in K$ if and only if for all $\sigma \in G$, $\sigma(x^m) = x^m$. By the definition of θ, this holds if and only if for all $\sigma \in G$, $\theta(\sigma)^m = 1$. So $|G| = [L : K]$ divides m. So $[L : K]$ must be the least m such that $x^m \in K$, as required. □

Corollary. Let K be a field that contains a primitive nth root of unity $\zeta = \zeta_n$. Let $a \in K^\times$. Then $f = T^n - a$ is irreducible over K if and only if a is not a dth power in K for any $1 \neq d \mid n$.

Proof. Let L be a splitting field for $f = T^n - a$, so $L = K(x)$ for $x^n = a$. Then the minimal polynomial of x divides f. So f is irreducible if and only if $f = m_{x,K}$, or equivalently, $[L : K] = n$.

Suppose $n = md$ for $d \neq 1$. Then a is a dth power in K if and only if $x^m \in K$ since $\zeta_n \in K$. By the above theorem, this holds if and only if $|G| \mid m$. □

Remark. This does not hold if we relax the assumption $\zeta_n \in K$. For example, consider $K = \mathbb{Q}$ and $T^4 + 4$.

Definition. Extensions of the form $L = K(x)$ where $x^n = a \in K$ and $\zeta_n \in K$ are called *Kummer extensions*.

Example. Let $n = 2$ and $\operatorname{char} K \neq 2$. Then $\zeta_2 = -1 \in K$. Then $K(\sqrt{a})/K$ is a quadratic Kummer extension if $a \notin (K^\times)^2$. Conversely, any quadratic extension must be of this form.

Theorem. Let K be a field that contains a primitive nth root of unity $\zeta = \zeta_n$ where $n > 1$. Let L/K be a Galois extension with cyclic Galois group of order n. Then L is a Kummer extension of K.

Proof. Let $\mathrm{Gal}(L/K) = \left\{1, \sigma, \sigma^2, \dots, \sigma^{n-1}\right\}$. For $y \in L$, let

$$x = R(y) = y + \zeta^{-1}\sigma(y) + \zeta^{-2}\sigma^2(y) + \dots + \zeta^{-(n-1)}\sigma^{n-1}(y) = \sum_{j=0}^{n-1} \zeta^{-j}\sigma^j(y) \in L$$

This is known as a *Lagrange resolvent*. Then

$$\sigma(x) = \sum_{j=0}^{n-1} \zeta^{-j}\sigma^{j+1}(y) = \sum_{j=0}^{n} \zeta^{1-j}\sigma^j(y) = \zeta x$$

Hence $\sigma(x^n) = \zeta^n x^n = x^n$, so $x^n \in K$. By the linear independence of field embeddings with $\{\sigma_i\} = \left\{1, \sigma, \dots, \sigma^{n-1}\right\}$, there exists y such that $R(y) = x \neq 0$. Now, since $\sigma^i x = \zeta^i x$, the $\sigma^i(x)$ are distinct, and so $\deg_K x = n$. In particular, $[K(x) : K] = n = [L : K]$, so $L = K(x)$. □

Example. Let $L/\mathbb{Q}$ be a Galois extension of degree 3. Since $\zeta_3 \notin \mathbb{Q}$, this is not a Kummer extension.

7. Trace and norm

7.1. Trace and norm

Let L/K be an extension of degree n, so L is a K-vector space of dimension n. Let $x \in L$. Then the map $U_x : L \to L$ defined by $U_x(y) = xy$ is K-linear, as it is L-linear. Since it is a linear map, it has a characteristic polynomial, a determinant, and a trace.

Definition. The *trace* and *norm* of $x \in L$ (relative to the extension L/K) are $\operatorname{Tr}_{L/K}(x) = \operatorname{tr} U_x \in K$ and $N_{L/K}(x) = \det U_x \in K$ respectively. The *characteristic polynomial* of $x \in L$ is $f_{x,L/K} = \det(TI - U_x) \in K[T]$ where I is the identity linear transformation.

We sometimes write tr_K, $\det_K$. Let $e_1, \dots, e_n$ be a basis for L/K. Then U_x can be written as a unique K-valued matrix $A = (a_{ij})$, so $xe_i = \sum_j a_{ji}e_j$. Then $\operatorname{Tr}_{L/K}(x) = \operatorname{tr}(A)$, and so on.

Example. Consider the quadratic extension $\mathbb{Q}(\sqrt{d})/\mathbb{Q}$ with the basis $1, \sqrt{d}$. Let $x = a+b\sqrt{d}$. Since $x \cdot 1 = a + b\sqrt{d}$ and $x \cdot \sqrt{d} = bd + a\sqrt{d}$,

$$A = \begin{pmatrix} a & bd \\ b & a \end{pmatrix}$$

Hence $\operatorname{Tr}_{L/K}(x) = 2a$ and $N_{L/K}(x) = a^2 - b^2d$.

Example. Consider $\mathbb{C}/\mathbb{R}$ with the basis $1, i$. Then the matrix of U_{x+iy} is

$$\begin{pmatrix} x & -y \\ y & x \end{pmatrix}$$

which is the usual encoding of complex numbers as 2 × 2 real matrices. Note the similarity between this matrix and the Cauchy–Riemann equations

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}; \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$$

Lemma. Let $x, y \in L$ and $a \in K$, where $n = [L : K]$. Then,

(i) $\operatorname{Tr}_{L/K}(x + y) = \operatorname{Tr}_{L/K}(x) + \operatorname{Tr}_{L/K}(y)$;

(ii) $N_{L/K}(xy) = N_{L/K}(x)N_{L/K}(y)$;

(iii) $N_{L/K}(x) = 0$ if and only if $x = 0$;

(iv) $\operatorname{Tr}_{L/K}(1) = n$ and $N_{L/K}(1) = 1$;

(v) $\operatorname{Tr}_{L/K}(ax) = a \operatorname{Tr}_{L/K}(x)$ and $N_{L/K}(ax) = a^n N_{L/K}(x)$.

In particular, $\operatorname{Tr}_{L/K}$ is K-linear and $N_{L/K} : L^\times \to K^\times$ is a homomorphism.

Proof. For part (iii), $N_{L/K}(x) = \det(U_x) \neq 0$ if and only if U_x is invertible. But this holds if and only if x is nonzero because L is a field. The other results follow from the laws of linear transformations. □

7.2. Formulae and applications

Theorem. Let $M/L/K$ be a tower of finite extensions. Then, for all $x \in M$,

$$\mathrm{Tr}_{L/K}(\mathrm{Tr}_{M/L}(x)) = \mathrm{Tr}_{M/K}(x); \quad N_{L/K}(N_{M/L}(x)) = N_{M/K}(x)$$

Proof. We prove the theorem for the trace; we will not need the result for the norm. Let $x \in M$. Let $u_1, \dots, u_m$ be a basis for M/L, and let $v_1, \dots, v_n$ be a basis for L/K. Let (a_{ij}) be the matrix of $U_{x,M/L}$, so $(a_{ij}) \in \mathrm{Mat}_{m,m}(L)$. Then $\mathrm{Tr}_{M/L}(x) = \sum_{i=1}^m a_{ii}$. For each (i,j), let the matrix of $U_{a_{ij}}$ be $A_{ij} \in \mathrm{Mat}_{n,n}(K)$. Then, $\mathrm{Tr}_{L/K}(\mathrm{Tr}_{M/L}(x)) = \sum_{i=1}^m \mathrm{Tr}_{L/K}(a_{ii}) = \sum_{i=1}^m \mathrm{tr}(A_{ii})$. Consider the basis $u_1v_1, \dots, u_1v_m, u_2v_1, \dots, u_nv_m$ for M over K. Then the matrix of $U_{x,M/K}$ is the block matrix

$$\begin{pmatrix} A_{11} & & & \\ & A_{22} & & \\ & & \ddots & \\ & & & A_{nn} \end{pmatrix}$$

which has trace $\sum_{i=1}^m \mathrm{tr}(A_{ii})$ as required. □

Proposition. Let $L = K(x)$, and $f = T^n + c_{n-1}T^{n-1} + \dots + c_0 \in K[T]$ be the minimal polynomial for x over K. Then $f_{x,L/K} = f$. Further, $\mathrm{Tr}_{L/K}(x) = -c_{n-1}$ and $N_{L/K}(x) = (-1)^n c_0$.

Proof. It suffices to prove the first statement, since the second follows from the fact that the determinant and trace are the given coefficients of the characteristic polynomial for any linear transformation. Consider the basis $1, x, \dots, x^{n-1}$ for L/K. Then, the matrix of U_x is

$$\begin{pmatrix} 0 & \cdots & & & & -c_0 \\ 1 & 0 & \cdots & & & -c_1 \\ 0 & 1 & 0 & \cdots & & \\ \vdots & 0 & 1 & 0 & \cdots & \\ & \vdots & 0 & 1 & \cdots & \\ & & \vdots & \vdots & \ddots & \\ & & & & & -c_{n-1} \end{pmatrix}$$

which has characteristic polynomial f since it is in rational canonical form. □

Corollary. Let $\mathrm{char}\, K = p > 0$, and $L = K(x)$ where $x \notin K$ but $x^p \in K$. Then for all $y \in L$, we have $\mathrm{Tr}_{L/K}(y) = 0$ and $N_{L/K}(y) = y^p$.

Proof. Recall that the minimal polynomial of x is $T^p - x^p$, so $[L : K] = p$. Suppose that $y \in K$. By a previous lemma, $\mathrm{Tr}_{L/K}(y) = py = 0$ and $N_{L/K}(y) = y^p$. Otherwise, since $[L : K]$ is prime, $K(y) = L$, and in particular, if $y = \sum a_i x^i$ then $y^p = \left(\sum a_i x^i\right)^p = \sum a_i (x^p)^i \in K$. So the minimal polynomial of y is $T^p - y^p$. Applying the previous proposition, the result follows. □

Proposition. Let L/K be a finite separable extension of degree n. Let $\sigma_1, \dots, \sigma_n : L \to M$ be the distinct K-homomorphisms of L into a normal closure M for L/K. Then

$$\mathrm{Tr}_{L/K}(x) = \sum_{i=1}^n \sigma_i(x); \quad N_{L/K}(x) = \prod_{i=1}^n \sigma_i(x); \quad f_{x,L/K} = \prod_{i=1}^n (T - \sigma_i(x))$$

Remark. If L/K is finite and Galois, then $\mathrm{Tr}_{L/K}(x) = \sum_{\sigma \in \mathrm{Gal}(L/K)} \sigma(x)$, and the other results are similar.

Proof. It suffices to show the result for the characteristic polynomial. Let $e_1, \dots, e_n$ be a basis for L/K. Let $P = (\sigma_i(e_j)) \in \mathrm{Mat}_{n,n}(M)$. Recall that the σ_i are linearly independent, so there exist no $y_i \in M$ such that for all j, $\sigma_i(e_j) = 0$. Hence P is nonsingular. Let $A = (a_{ij})$ be the matrix of U_x, so $xe_j = \sum_r a_{rj} e_r$. Applying σ_i, we have

$$\sigma_i(x)\sigma_i(e_j) = \sum_r \sigma_i(e_r) a_{rj}$$

So if S is the diagonal matrix with (i,i)th entry $\sigma_i(x)$, then the given equation can be rewritten as $SP = PA$. Therefore $S = PAP^{-1}$. So S and A are conjugate matrices and hence have the same characteristic polynomial. We explicitly find that the characteristic polynomial of S is $\prod(T - \sigma_i(x))$ and the characteristic polynomial of A is $f_{x,L/K}$. So they are equal as required. □

Note that since the trace $\mathrm{Tr}_{L/K} : L \to K$ is K-linear, it is either the zero map or surjective.

Theorem. Let L/K be a finite extension. Then, L/K is separable if and only if $\mathrm{Tr}_{L/K}$ is surjective.

Remark. If $\mathrm{char}\, K = 0$, $\mathrm{Tr}_{L/K}(1) = n \neq 0$, so the result holds easily.

Proof. Suppose L/K is separable, and $\sigma_i, \dots, \sigma_n$ are the K-homomorphisms of L into a normal closure M of L/K. Then $\mathrm{Tr}_{L/K}(x) = \sum_{i=1}^n \sigma_i(x)$. As the σ_i are linearly independent, there exists x such that $\sum_{i=1}^n \sigma_i(x) \neq 0$. So $\mathrm{Tr}_{L/K}(x) \neq 0$, and in particular, it must be surjective as it is K-linear.

Now suppose L/K is inseparable. Then there exists $x \in L$ such that $K(x) \supsetneq K(x^p)$ from example 7 on example sheet 2. As we have shown, $\mathrm{Tr}_{K(x)/K(x^p)} = 0$, so

$$\mathrm{Tr}_{L/K} = \mathrm{Tr}_{L/K(x)} \circ \mathrm{Tr}_{K(x)/K(x^p)} \circ \mathrm{Tr}_{K(x^p)/K} = 0$$

□

Example. Consider the extension of finite fields $\mathbb{F}_{q^n}/\mathbb{F}_q$ for $q = p^r$. This is separable, so there exists $x \in \mathbb{F}_{q^n}$ such that $\mathrm{Tr}(x) = 1$. It is also possible to prove this directly by using the fact that the multiplicative group is cyclic.

Remark. This criterion can be used to give another proof that if M/L and L/K are separable, M/K is also separable.

8. Algebraic closure

8.1. Definition

Definition. A field K is *algebraically closed* if every non-constant polynomial over K splits into linear factors over K.

Remark. An equivalent condition is that the only irreducible polynomials are linear.

Example. The complex numbers $\mathbb{C}$ form an algebraically closed field due to the fundamental theorem of algebra.

Proposition. The following are equivalent.

(i) K is algebraically closed.

(ii) If L/K is a field extension and $x \in L$ is algebraic over K, then $x \in K$.

(iii) If L/K is an algebraic extension, $L = K$.

Proof. (i) implies (ii). Let L/K be a field extension and $x \in L$ algebraic over K. Let f be the minimal polynomial for x over K. Then f is linear, so $x \in K$.

(ii) implies (iii). An extension L/K is algebraic when all $x \in L$ are algebraic over K. So $x \in K$ by (ii).

(iii) implies (i). Let f be an irreducible polynomial, and $L = {}^{K[T]}\!/_{(f)}$, so L/K is a finite algebraic extension. Then $L = K$, so f is linear. □

Proposition. Let L/K be an algebraic extension such that every irreducible polynomial $f \in K[T]$ splits into linear factors in L. Then L is algebraically closed.

Such a field is called an *algebraic closure* of K.

Proof. Let M/L be an extension, and let $x \in M$ be algebraic over L. Then x is algebraic over K. By hypothesis, its minimal polynomial $m_{x,K} \in K[T]$ splits into linear factors over L. So $x \in L$. By criterion (ii) in the previous proposition, L is algebraically closed. □

Remark. An algebraic closure of K is the same as an algebraic extension of K which is algebraically closed.

Corollary. The field $\overline{\mathbb{Q}}$ of algebraic complex numbers is algebraically closed. In particular, $\overline{\mathbb{Q}}$ is an algebraic closure of $\mathbb{Q}$.

Proof. We apply the previous result to the extension $\overline{\mathbb{Q}}/\mathbb{Q}$. The extension is algebraic, so it suffices to check that every irreducible polynomial $f \in \mathbb{Q}[T]$ splits into linear factors in $\overline{\mathbb{Q}}$. By the fundamental theorem of algebra, f splits in $\mathbb{C}$. By definition of $\overline{\mathbb{Q}}$, we have $f = \prod(T - x_i)$ where each $x_i \in \overline{\mathbb{Q}}$ as required. □

8.2. Algebraic closures of countable fields

Proposition. Let K be a countable field. Then K has an algebraic closure.

Proof. If K is a countable field, then $K[T]$ is a countable ring. We will enumerate the monic irreducible polynomials $f_i \in K[T]$ for $i \geq 1$. Let $L_0 = K$, and inductively define L_i to be a splitting field for f_i over L_{i-1}.

One can perform this in such a way that no choices need to be made in the construction of the splitting fields. We may also assume that $L_{i-1} \subseteq L_i$ for each $i \geq 1$, because if $\sigma : L_{i-1} \to L_i$ is the extension, we can replace L_i with $L_{i-1} \sqcup (L_i \setminus \sigma(L_{i-1}))$. Let $L = \bigcup L_i$ be their union. By construction, every f_i splits in L, so L is an algebraic closure of K. □

Example. $\mathbb{F}_p$ has an algebraic closure.

8.3. Zorn's lemma

For a general field, we need to apply some set-theoretic machinery.

Definition. A binary relation $\preceq$ on a set S is a *partial order* if it is reflexive, transitive, and antisymmetric. Explicitly, for all $x, y, z \in S$, we have

$$x \preceq x; \quad x \preceq y, y \preceq z \implies x \preceq z; \quad x \preceq y, y \preceq x \implies z = y$$

We say $(S, \preceq)$ is a *partially ordered set*, or a *poset*. It is *totally ordered* if the order is total; $x \preceq y$ or $y \preceq x$ for all $x, y \in S$.

Definition. Let S be a partially ordered set. A *chain* in S is a totally ordered subset. An *upper bound* for a subset T of S is an element $z \in S$ such that for all $x \in T$, we have $x \preceq z$. A *maximal element* of S is an element $y \in S$ such that for all $x \in S$, $y \preceq x$ implies $y = x$.

If S is totally ordered, S has at most one maximal element.

Lemma (Zorn)**.** Let S be a nonempty partially ordered set. Suppose that every chain in S has an upper bound in S. Then S has a maximal element.

This can be proven using the axiom of choice.

Example. Let V be a vector space over K. Then V has a basis; a set $B \subseteq V$ such that any finite subset of B is linearly independent, and for all $v \in V$, there exists $b_1, \dots, b_k \in B$ and $a_1, \dots, a_k \in K$ such that $v = \sum_{i=1}^k a_i b_i$. If $V = \{0\}$, the result is trivial by taking $V = \varnothing$. Otherwise, let S be the set of all subsets $X \subseteq V$ where finite subsets of X are linearly independent. S is ordered by inclusion; this is a partial order. S is nonempty since $V \neq \{0\}$. Each chain $T \subseteq S$ has an upper bound by taking its union $Y = \bigcup_{X \in T} X$. This upper bound indeed lies in S, since we only need to check finite subsets of Y for linear independence. Then by Zorn's lemma, S has a maximal element B, which can be seen to be a basis.

Proposition. Let L/K be an algebraic extension, and let M be algebraically closed. Let $\sigma : K \to M$. Then there exists $\overline{\sigma} : L \to M$ extending σ.

Proof. First, consider the case $L = K(x)$ where x is algebraic over K with minimal polynomial $m_{x,K} = f$. Then $\sigma f \in M[T]$. Since M is algebraically closed, σf splits in M. Therefore there exists such a $\overline{\sigma} : K(x) \to M$ extending σ. We can obtain one homomorphism for each root of σf in M.

Now consider the general case. Suppose $K \subseteq L$ without loss of generality, by replacing K with its image in L. Let

$$S = \left\{ (F, \tau) \,\middle|\, K \subseteq F \subseteq L, \tau : F \to M, \tau\Big|_K = \sigma \right\}$$

This has a partial order given by $(F, \tau) \preceq (F', \tau')$ where $F \subseteq F'$ and $\tau'|_F = \tau$. Therefore, S is a partially ordered set. It contains (K, σ), so it is not empty.

Let $T = (F_i, \tau_i)_{i \in I}$ be a chain in S. If T is empty, we can vacuously upper bound it with (K, σ). Otherwise, we define $F' = \bigcup_{i \in I} F_i$. This is a field since T is a chain; in particular, for all $i, j \in I$, we have either $F_i \subseteq F_j$ or $F_j \subseteq F_i$. Now define $\tau' : F' \to M$ by mapping x to $\tau_i(x)$ where $x \in \tau_i$; this is independent of the choice of i since $\tau_j\Big|_{F_i} = \tau_i$ and T is a chain. This is an upper bound in S for the chain.

Then, by Zorn's lemma, S has a maximal element. Let (F, τ) be this maximal element. We will show $F = L$; in this case, $\tau = \overline{\sigma}$ is an extension as required.

Clearly $F \subseteq L$. If $x \in L$, then by the first part applied to $F(x)/F$, we can extend the homomorphism $\tau : F \to M$ into a homomorphism $\overline{\tau} : F(x) \to M$. Then $(F(x), \overline{\tau}) \in S$, and $(F, \tau) \preceq (F(x), \overline{\tau})$. By maximality, $F(x) = F$, so $x \in F$. Hence $F = L$ as required. □

8.4. Algebraic closures of general fields

One can construct an algebraic closure of a field using Zorn's lemma, obtaining a field that extends all algebraic extensions of a given field. However, difficulties arise since the class of algebraic extensions of a field does not form a set. Zorn's lemma can be utilised inside a suitably well-behaved set, but instead, we will construct the algebraic closure via the maximal ideal theorem.

Theorem (maximal ideal theorem). Let R be a non-zero commutative ring with a 1. Then R has a maximal ideal.

Proof sketch. Let S be the set of all proper ideals $I \lhd R$, partially ordered by inclusion. A maximal ideal is a maximal element of S. We apply Zorn's lemma. Let T be a nonempty chain, since anything is an upper bound for an empty chain. Then $J = \bigcup_{I \in T} I$ is an ideal. As $1 \notin I$ for all $I \in T$, we conclude $1 \notin J$. So J is a proper ideal, and hence is an upper bound. □

Theorem. Let K be a field. Then K has an algebraic closure $\overline{K}$. If $\sigma : K \to K'$ is an isomorphism, and $\overline{K}, \overline{K}'$ are any algebraic closures of K, K', then σ extends to an isomorphism $\overline{\sigma} : \overline{K} \to \overline{K}'$.

Remark. The extension $\overline{\sigma}$ is not generally unique.

Proof. We begin by proving the existence of the algebraic closure. Let P be the set of monic irreducible polynomials in $K[T]$, and construct K_1 such that every $f \in P$ has a root in K_1. First, we will find a ring in which every $f \in P$ has a root.

Let $R = K\left[\{T_f\}_{f\in P}\right]$ be the set of finite K-linear combinations of monomials $T_{f_1}^{m_1} \dots T_{f_k}^{m_k}$ for $f_i \in P$. Let I be the ideal generated by $f(T_f)$ for each $f \in P$. Now, in ${}^{R}\!/_{I}$, $T_f + I$ is a root of f.

We must check that $I \neq R$. If $I = R$, then in particular $1 \in I$. In other words, for some finite subset $Q \subseteq P$, there exists $r_f \in R$ for $f \in Q$ such that $1 = \sum_{f\in Q} r_f f(T_f)$. Enlarging Q if necessary, we can assume that each r_f is a polynomial in $\{T_g \mid g \in Q\}$. Let L/K be a splitting field for $\prod_{f\in Q} f$, and $a_f \in L$ be a root of f for each $f \in Q$. Consider the homomorphism $\varphi : R \to L$ such that $\varphi|_K = \mathrm{id}$ and $\varphi(T_f) = a_f$ for $f \in Q$, and $\varphi(T_f) = 0$ for $f \notin Q$. Then

$$1 = \varphi(1) = \sum_{f\in Q} \varphi(r_f f(T_f)) = \sum_{f\in Q} \varphi(r_f) f(a_f) = 0$$

This is a contradiction, so I is in fact a proper ideal.

By the maximal ideal theorem, the ring ${}^{R}\!/_{I}$ has a maximal ideal $\overline{J}$. Equivalently, there exists a maximal ideal J of R containing I, since the ideals of ${}^{R}\!/_{I}$ are in bijection with the ideals of R containing I by the isomorphism theorem. Now let $K_1 = {}^{R}\!/_{J}$. This is a field since J is maximal. Let $x_f = T_f + J \in K_1$, then K_1/K is generated by the x_f, and $f(x_f) = 0$ by construction. So K_1/K is an algebraic extension of K in which every $f \in P$ has a root.

Let P_1 be the set of monic irreducibles in $K_1[T]$. We apply the same procedure to K_1 and P_1 to obtain a field K_2, and so on. We then obtain a tower $K \subseteq K_1 \subseteq K_2 \subseteq \cdots$ such that if $f \in K_n[T]$ is non-constant, it has a root in K_{n+1}.

Now, suppose $f \in K[T]$ is non-constant. Then we can write $f = (T - x_1)f_1$ where $x_1 \in K_1, f_1 \in K_1[T]$, and so on. So f splits in $K_{\deg f - 1}$. Therefore, the union $\bigcup_{n\in\mathbb{N}} K_n$ is algebraically closed, and hence is an algebraic closure of K.

We now prove uniqueness. Let $K \subseteq \overline{K}$ and $K' \subseteq \overline{K}'$ be algebraic closures, and let $\sigma : K \to K'$ be an isomorphism. Then by the previous result, as $\overline{K}/K$ is algebraic, σ extends to a homomorphism $\overline{\sigma} : \overline{K} \to \overline{K}'$. It suffices to show that σ is an isomorphism. We have $K' \subseteq \sigma(\overline{K}) \subseteq \overline{K}'$, so $\overline{K}'/\sigma(\overline{K})$ is algebraic. $\overline{K}$ is algebraically closed, so $\sigma(\overline{K})$ is also algebraically closed. So $\overline{K}' = \sigma(\overline{K})$ by part (iii) of a previous result. □

9. Solving polynomial equations

9.1. Cubics

Let $f \in K[T]$ be a monic separable cubic. Then $G = \mathrm{Gal}(f/K) \leq S_3$ acting on the roots x_1, x_2, x_3 in a splitting field L of K. If f is reducible, f is either a product of three linear factors, in which case G is trivial, or f is a linear factor multiplied by a quadratic, in which case G is isomorphic to S_2.

Now suppose f is irreducible. We will assume that $\operatorname{char} K \neq 2, 3$. We have $G = S_3$ or $G = A_3$. We know that $G = A_3$ if and only if the discriminant $\mathrm{Disc}(f)$ is a square in $K^\times$. In general, the Galois correspondence yields

$$\begin{tikzcd}
L = K(x_1, x_2, x_3) \arrow[d, "3 \text{ if } f \text{ irreducible, else } 1"] & \{1\} \\
K_1 = K(\Delta) = L^{G \cap A_3} \arrow[d, "2 \text{ or } 1"] & G \cap A_3 \arrow[u] \\
K & G \arrow[u]
\end{tikzcd}$$

Then $K_1 = K(\sqrt{\mathrm{Disc}(f)})$, and $K_1 = L$ if f is reducible.

In the irreducible case, L/K_1 is Galois with $\mathrm{Gal}(L/K_1) \simeq {}^{\mathbb{Z}}\!/_{3\mathbb{Z}}$. Recall that if $\omega \in K_1$ is a primitive third root of unity, then $L = K_1(y)$ where $y^3 \in K_1$, by Kummer theory.

We can compute this y explicitly. Suppose $f = T^3 + bT + c$ without loss of generality. Then $\Delta^2 = -4b^3 - 27c^2$. If $b = 0$, the roots of f are $w^i\sqrt[3]{-c}$, so let y be any of them. In the other case $b \neq 0$, let y be a Lagrange resolvent. If the roots of f in L are x_1, x_2, x_3, take $y = x_1 + \omega^2 x_2 + \omega x_3 = (1-\omega)(x_1 - \omega x_2)$ as $x_1 + x_2 + x_3 = 0$. Then $L(\omega) = K(\Delta, \omega, y)$ if and only if $y \neq 0$, by the proof of the structure of Kummer extensions. Let $y' = x_1 + \omega x_2 + \omega^2 x_3$, then $yy' = -3b \neq 0$ since we are not in characteristic 3. Note that $y + y' = y + y' + x_1 + x_2 + x_3 = 3x_1$. One can calculate $y^3 = \frac{1}{2}\left(-3\sqrt{-3}\Delta + 27c\right)$, so $x_1 = y - \frac{3b}{y}$.

If not, let $L(\omega)$ be the splitting field of $f \cdot (T^3 - 1)$ over K. Then $L(\omega)/K_1(\omega)$ is Galois with Galois group ${}^{\mathbb{Z}}\!/_{3\mathbb{Z}}$ as before. So $L(\omega) = K_1(\omega, y)$ where $y^3 \in K_1(\omega)$.

Therefore, in every case, x_i lie in the field obtained by adjoining successive square roots and cube roots to K, since $\omega = \frac{-1+\sqrt{-3}}{2}$. This is a theoretical description of Cardano's solution to the cubic.

9.2. Quartics

Let $f \in K[T]$ be a monic separable quartic, with $\operatorname{char} K \neq 2, 3$. Then $\mathrm{Gal}(f/K) \leq S_4$. Note that S_4 acts on the partitions $(12 \mid 34), (13 \mid 24), (14 \mid 23)$ of $\{1, 2, 3, 4\}$. Then we have a homomorphism $S_4 \to S_3$. The kernel of this homomorphism is the Klein four-group $V =$

$\{e, (12)(34), (13)(24), (14)(23)\} \lhd S_4$. Hence the homomorphism is surjective, as $|V| \cdot |S_3| = |S_4|$.

Let f have splitting field L with (distinct) roots $x_1, \dots, x_4$. Suppose that $x_1 + \dots + x_4 = 0$ without loss of generality as the characteristic is not 2, so $f = T^4 + aT^2 + bT + c$. Since V is a normal subgroup of S_4, $G \cap V$ is a normal subgroup of G and contains V. In particular, we have a homomorphism ${}^{G}\!/_{G \cap V} \hookrightarrow {}^{S_4}\!/_{V} \simeq S_3$. But ${}^{G}\!/_{G \cap V} = \mathrm{Gal}(M/K)$. So we should be able to write M as the splitting field of a cubic $g \in K[T]$.

Let $y_{12} = x_1 + x_2 = -(x_3 + x_4) = -y_{34}$, and let $y_{13}, y_{24}, y_{14}, y_{23}$ be defined similarly. Note that $G \cap V$ maps y_{12} to y_{12} or $y_{34} = -y_{12}$, and so on. So $y_{12}^2, y_{13}^2, y_{14}^2$ are fixed under $G \cap V$. Hence they lie in $M = L^{G \cap V}$.

Suppose $y_{12}^2 = y_{13}^2$. Then either $y_{12} = y_{13}$, so $x_2 = x_3$, contradicting separability, or $y_{12} = -y_{13}$, so $2x_1 + x_2 + x_3 = 0$, giving $x_1 = x_4$, also contradicting separability. So these are distinct elements of M, and hence are indeed the roots of a separable cubic $g \in K[T]$. This is called the *resolvent cubic*.

$M = L^{G \cap V}$ is a splitting field of g. Note that $x_1 = \frac{1}{2}(y_{12} + y_{13} + y_{14})$ and similar results hold for x_2, x_3, x_4. Hence $L = M(y_{12}, y_{13}, y_{14})$. We can compute $g = (T - y_{12}^2)(T - y_{13}^2)(T - y_{14}^2) = T^3 + 2aT^2 + (a^2 - 4c)T - b^2$. In particular, $y_{12}y_{13}y_{14} = b$, hence we can simplify to $L = M(y_{12}, y_{13})$ where $y_{12}^2, y_{13}^2 \in M$.

In conclusion, we have found a way to solve $f = 0$. First, we solve the resolvent equation $g = 0$, and then we take at most two square roots to obtain the relevant field generators.

9.3. Solubility by radicals

Let $f \in K[T]$ be a monic polynomial in a field K of characteristic zero. To prove that there is no quintic formula, we must first establish a definition of 'formula'. The relevant notion is solubility by radicals.

Definition. An irreducible polynomial $f \in K[T]$ is *soluble by radicals* over K if there exists a sequence of fields $K = K_0 \subseteq K_1 \subseteq \dots \subseteq K_m$, with $x \in K_m$ a root of f, and each K_i is obtained from K_{i-1} by adjoining a root, so $K_i = K_{i-1}(y_i)$ where $y_i^{d_i} \in K_{i-1}$.

Remark. This is a generalisation of ruler and compass constructions to permit roots of arbitrary degree.

Note that we can adjoin extra roots if desired. In particular, adjoining roots of unity, f is soluble by radicals over K if there exists $d \geq 1$ and $K = K_0 \subseteq \dots \subseteq K_m$, such that $x \in K_m$ is a root of f, and $K_1 = K_0(\zeta_d)$ where ζ_d is a primitive dth root of unity. We can also assume that the other extensions satisfy $K_i = K_{i-1}(y_i)$ for $y_i^d = a_i \in K_{i-1}$. This condition can be easily satisfied by letting d be the least common multiple of the d_i that occurs in the tower of fields.

Note that K_1/K_0 is Galois with abelian Galois group. K_i/K_{i-1} for $i > 1$ is Galois, where the Galois group is a subgroup of ${}^{\mathbb{Z}}\!/\!{}_{d\mathbb{Z}}$ as it is a Kummer extension.

To obtain all roots of f, we consider a normal closure M of K_m; this will contain a splitting field for f, since it contains one root and f is irreducible. To determine M, let $K_i' \subseteq M$ be a normal closure of K_i for each i. As we are in characteristic zero, an extension is Galois if and only if it is normal. Note that K_1 is Galois, so $K_1 = K_1' = K(\zeta_d)$.

Proposition. $K_i' = K_{i-1}'\left(\left\{\sqrt[d]{\sigma(a_i)} \,\middle|\, \sigma \in \mathrm{Gal}(K_{i-1}'/K)\right\}\right)$.

Proof. Suppose $\sigma \in \mathrm{Gal}(K_{i-1}'/K)$. Then we can lift σ to an element $\overline{\sigma} \in \mathrm{Gal}(K_i'/K)$ such that $\overline{\sigma}|_{K_{i-1}'} = \sigma$. Since K_i'/K is normal, it contains $\overline{\sigma}(y_i)$, and $\overline{\sigma}(y)^d = \sigma(y^d) = \sigma(a_i)$. So the right hand side is contained in K_i'.

It suffices to show the right hand side is a normal extension. It is the splitting field over K_{i-1}' of the polynomial $g_i = \prod_{\sigma\in\mathrm{Gal}(K_{i-1}'/K)}(T^d - \sigma(a_i))$. This has coefficients in K. If K_{i-1}' is the splitting field of some polynomial h_{i-1} over K, then the right hand side is the splitting field of the product $g_i h_{i-1}$ over K. So it is normal. □

Proposition. $\mathrm{Gal}(K_i'/K_{i-1}')$ is abelian.

Proof. This proof is a variant on the proof of a previous theorem. Consider the case $i > 1$. Let $A = \mathrm{Gal}(K_i'/K_{i-1}')$. Let $\tau \in A$ and $\sigma \in \mathrm{Gal}(K_i'/K)$. Then $\tau(\sqrt[d]{\sigma(a_i)}) = \zeta_d^{m_\sigma}\sqrt[d]{\sigma(a_i)}$ where $m_\sigma \in {}^{\mathbb{Z}}\!/\!{}_{d\mathbb{Z}}$. Hence $\tau \mapsto (m_\sigma) \in \left({}^{\mathbb{Z}}\!/\!{}_{d\mathbb{Z}}\right)^r$ is an injective homomorphism, where $r = |\mathrm{Gal}(K_{i-1}'/K)|$.

If $i = 1$, then $K_1 = K(\zeta_d)$. So the Galois group is a subgroup of $\left({}^{\mathbb{Z}}\!/\!{}_{d\mathbb{Z}}\right)^\times$, so is abelian. □

Since all of the fields K_i' are normal closures, the N_i are normal subgroups of G.

Definition. A finite group G is *soluble* if there exists a chain of normal subgroups $N_i \trianglelefteq G$ with $G = N_0 \supseteq N_1 \supseteq \cdots \supseteq N_m = \{1\}$ such that ${}^{N_i}\!/\!{}_{N_{i+1}}$ is abelian for all i.

Example. Any abelian group is soluble. S_3 is soluble, by considering the chain $S_3 \supset A_3 \supset \{1\}$, as ${}^{S_3}\!/\!{}_{A_3} \simeq {}^{\mathbb{Z}}\!/\!{}_{2\mathbb{Z}}$ and $A_3 \simeq {}^{\mathbb{Z}}\!/\!{}_{3\mathbb{Z}}$. S_4 is also soluble; the chain $S_4 \supset A_4 \supset V \supset \{1\}$ suffices. Note that ${}^{S_4}\!/\!{}_{A_4} \simeq {}^{\mathbb{Z}}\!/\!{}_{2\mathbb{Z}}, {}^{A_4}\!/\!{}_{V} \simeq {}^{\mathbb{Z}}\!/\!{}_{3\mathbb{Z}}, V \simeq \left({}^{\mathbb{Z}}\!/\!{}_{2\mathbb{Z}}\right)^2$.

We have shown that ${}^{N_i}\!/\!{}_{N_{i+1}} = \mathrm{Gal}(K_i'/K_{i-1}')$ is abelian. Hence $\mathrm{Gal}(M/K)$ is soluble.

Lemma. Every subgroup and quotient of a soluble group is soluble.

Proof. Let $G = N_0 \supset N_1 \supset \dots N_m = \{1\}$, where the quotients ${}^{N_i}\!/\!{}_{N_{i+1}}$ are abelian. Let $H \leq G$. Then $H \cap N_i \trianglelefteq H$, and there is an injective homomorphism from ${}^{H \cap N_i}\!/\!{}_{H \cap N_{i+1}}$ to ${}^{N_i}\!/\!{}_{N_{i+1}}$. Hence the ${}^{H \cap N_i}\!/\!{}_{H \cap N_{i+1}}$ are abelian, so H is soluble.

Now let $\pi : G \to \overline{G} = {}^{G}\!/_{H}$ for $H \trianglelefteq G$. Then $\pi(N_i) \trianglelefteq \overline{G}$, and ${}^{N_i}\!/_{N_{i+1}}$ surjects onto ${}^{\pi(N_i)}\!/_{\pi(N_{i+1})}$. □

Theorem (Abel–Ruffini)**.** Let $f \in K[T]$ be soluble by radicals over K. Then $\mathrm{Gal}(f/K)$ is soluble.

Proof. $\mathrm{Gal}(f/K) = \mathrm{Gal}(L/K) \simeq {}^{\mathrm{Gal}(M/K)}\!/_{\mathrm{Gal}(M/L)}$. We know that $\mathrm{Gal}(M/K)$ is soluble, so the result follows from the fact that quotients of soluble groups are soluble. □

Remark. One can easily show the converse to this theorem.

Proposition. If $n \geq 5$, then S_n and A_n are insoluble.

Proof. S_n and A_n contain A_5 as a subgroup, so it suffices to show that A_5 is insoluble. A_5 is not abelian, and it is simple, so it is insoluble. □

Corollary. Let $n = \deg f \geq 5$, and $A_n \leq \mathrm{Gal}(f/K)$. Then f is not soluble by radicals over K.

10. Miscellaneous results

10.1. Fundamental theorem of algebra

This subsection is non-examinable. We show that $\mathbb{C}$ is algebraically closed over $\mathbb{Q}$, without using complex analysis. We will only use the following facts:

(i) every polynomial of odd degree over $\mathbb{R}$ has a root, due to the intermediate value theorem;

(ii) every quadratic over $\mathbb{C}$ splits into linear factors, so we can take square roots;

(iii) every finite group G has a subgroup H such that $(G : H)$ is odd and $|H|$ is a power of 2, by Sylow's theorem for $p = 2$;

(iv) if G is a p-group, so $|G| = p^k$ and $k > 0$, then G has a subgroup of index p, since G has a non-trivial centre.

Let $K/\mathbb{C}$ be a finite extension. Let L/K be a normal closure of K over $\mathbb{R}$, so L is a Galois extension of $\mathbb{R}$ containing $\mathbb{C}$. Let $G = \mathrm{Gal}(L/\mathbb{R})$. We will show that $L = \mathbb{C}$.

Let $H \leq G$ be a Sylow 2-subgroup, and consider L^H. We have $[L^H : \mathbb{R}] = (G : H)$, which is odd. So if $x \in L^H$, by (i), its minimal polynomial is linear over $\mathbb{R}$, so $x \in \mathbb{R}$. Hence $L^H = \mathbb{R}$, so $H = G$. So G is a 2-group.

Let $G \supset G_1 = \mathrm{Gal}(L/\mathbb{C})$, and $G_2 \leq G_1$ be a subgroup of index 2, which exists by (iv). Then $[L^{G_2} : \mathbb{C}] = (G_1 : G_2)$, contradicting the fact (ii) that quadratics split in $\mathbb{C}$. So there cannot exist a subgroup of index 2, so $G_1 = \{e\}$, and $L = \mathbb{C}$.

10.2. Artin's theorem on invariants

Theorem (Artin). Let L be a field and $G \leq \mathrm{Aut}(L)$ be a finite subgroup of automorphisms of L. Define $L^G = \{x \in L \mid \forall \sigma \in G, \sigma(x) = x\}$. Then L/L^G is finite, and satisfies $[L : L^G] = |G|$.

Remark. Unlike in the Galois correspondence, this theorem does not rely on a field extension, just a single field and a finite group of automorphisms. In particular, we find that L/L^G is finite and Galois, with Galois group G.

Proof. It suffices to show L/L^G is finite, because then we can apply the Galois correspondence to show $[L : L^G] = |G|$. Let $K = L^G$, and let $x \in L$. Then if $\{\sigma_1(x), \dots, \sigma_r(x)\}$ is the orbit of G on x, then x is a root of $f = \prod_{i=1}^r (T - \sigma_i(x))$. But $f \in L^G[T] = K[T]$. By construction, f is separable. Hence x is algebraic and separable over K, and $\deg_K x \leq |G|$.

Let $y \in L$ have maximal degree. We claim that $K(y) = L$. If not, there exists $x \in L \setminus K(y)$. By above, x, y are algebraic and separable over K. By the primitive element theorem, there exists $z \in L$ such that $K(x, y) = K(z) \supsetneq K(y)$, so $\deg_K z > \deg_K y$. But y was chosen to have maximal degree, so this is a contradiction. □

Remark. One can prove this theorem directly without appealing to the Galois correspondence or the primitive element theorem. This can then be used as a starting point for Galois theory, which then allows the more complicated theorems to be proven.

There are two common ways to construct finite Galois extensions. The first, studied earlier in the course, involves taking the splitting field of a separable polynomial; this method constructs a larger field from a given base field. Artin's theorem provides another way to construct such extensions, by fixing a large field L and constructing the subfield L^G.

Example. Let $\Bbbk$ be a field, and let $L = \Bbbk(X_1, \dots, X_n)$ be the field of rational functions, defined as the fractions of the polynomial ring $\Bbbk[X_1, \dots, X_n]$. Let $G = S_n$ be the symmetric group permuting the X_i. Then $G \leq \mathrm{Aut}(L)$.

Theorem. Let $\Bbbk$ be a field and let $L = \Bbbk(X_1, \dots, X_n)$. Then $L^G = \Bbbk(s_1, \dots, s_n)$.

Proof. Recall that $\Bbbk[X_1, \dots, X_n]^G = \Bbbk[s_1, \dots, s_n]$ where the s_i are the elementary symmetric polynomials in the X_i, and there are no nontrivial relations between the s_i. In particular, $\Bbbk(s_1, \dots, s_n) \subseteq L^G$.

Conversely, let $\frac{f}{g} \in L^G$ for $f, g \in \Bbbk[X_1, \dots, X_n] = R$. Without loss of generality let f, g be coprime. Then for all $\sigma \in G$, $\frac{f}{g} = \frac{\sigma f}{\sigma g}$. By Gauss' lemma, R is a unique factorisation domain, and the units in R are the nonzero constants $\Bbbk^\times$. Hence $\sigma f = c_\sigma f$ and $\sigma g = c_\sigma g$ where $c_\sigma \in \Bbbk^\times$.

Since G is finite and has order $N = n!$, $f = \sigma^N f = c_\sigma^N f$. So c_σ is an Nth root of unity. Then fg^{N-1}, g^N are invariant under σ, so $fg^{N-1}, g^N \in R^G = \Bbbk[s_1, \dots, s_n]$. So $\frac{f}{g} = \frac{fg^{N-1}}{g^N} \in \Bbbk(s_1, \dots, s_n)$. □

Example. Let $L = \Bbbk(X_1, \dots, X_n)$, and let $K = \Bbbk(s_1, \dots, s_n) = L^G$ where $G = S^n$. Then by Artin's theorem, L/K is a finite Galois extension with Galois group G. Let $f = T^n - s_1 T^{n-1} + \dots + (-1)^n s_n \in K[T]$. Then in L, $f = \prod_{i=1}^n (T - X_i)$. Since the X_i are different, f is separable, and L/K is a splitting field for f. Hence $\mathrm{Gal}(f/K) = S^n$. Informally, the general polynomial of degree n has Galois group S^n. It is not difficult to show that for any finite group G, there exists a Galois extension with Galois group isomorphic to G.

10.3. Other areas of study

This is one of a number of theories in *invariant theory*, in which one considers a ring R and a group $G \leq \mathrm{Aut}(R)$, and study R^G. If R is a polynomial ring $\Bbbk[X_1, \dots, X_n]$ and $G \leq S_n$, then knowing R^G can help with the computation of Galois groups algorithmically. For example, if $G = A_n$, then $\Bbbk[X_1, \dots, X_n]^{A_n} = \Bbbk[s_1, \dots, s_n, \Delta]$ where $\Delta = \prod_{i<j}(X_i - X_j)$, for $\operatorname{char} \Bbbk \neq 2$.

Now consider $R = \Bbbk[X_1, X_2]$ and $G = \{1, \sigma\}$ where $\sigma(X_i) = -X_i$. Let $\operatorname{char} \Bbbk \neq 2$. Then one can show $R^G = \Bbbk[X_1^2, X_2^2, X_1 X_2] = {}^{\Bbbk[Y_1, Y_2, Y_3]}\!/\!_{(Y_1 Y_2 - Y_3^2)}$. Geometrically, $\{Y_1 Y_2 = Y_3^2\} \subset$

$\mathbb{R}^3$ is a double cone. The point at which the cones meet is known as a singularity; such singularities occur in the study of algebraic geometry.

If K and G are fixed, it is not always the case that there exists a Galois extension L/K such that $\text{Gal}(L/K) = G$. For instance, if K is algebraically closed, it has no nontrivial Galois extensions. If $K = \mathbb{F}_p$, then $\text{Gal}(L/K)$ must be cyclic.

The *inverse Galois problem* asks whether every finite group G is the Galois group of some Galois extension $L/\mathbb{Q}$. This is unsolved in the general case. On the extra example sheet, one shows that every abelian group is in fact the Galois group of some Galois extension $L/\mathbb{Q}$. There is a famous theorem by Shafarevich that every finite soluble group is such a Galois group over $\mathbb{Q}$. This is also known to hold for most finite simple groups; in particular, due to a theorem of John Thompson, the monster group is known to be a Galois group over $\mathbb{Q}$.

Perhaps to solve this problem, it would be better to instead understand $\text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. The inverse Galois problem is equivalent to asking whether every finite group is a quotient of $\text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. We may also be interested in finding the representations of this group. This leads to the Langlands programme.

VI. Coding and Cryptography

Lectured in Lent 2023 by Prof. S. Martin

Coding theory allows us to mathematically reason about methods of communication. The theory is largely broken into two parts: noiseless coding and noisy coding.

Noiseless codes describe ways to compress data into less space. Such schemes have been in use since at least the time of the Ancient Greeks, who used torches placed upon hilltops to concisely convey Greek letters over a long distance. We explore some examples of noiseless codes, and prove theoretical results about how efficiently they can be expected to encode data from a variety of sources.

Noisy coding aims to detect, or possibly correct, errors that may have been introduced while transmitting data across a noisy channel. An example of such a code is the International Standard Book Number: it detects any digit typed incorrectly, and any transposition of two adjacent digits. We investigate how reliably channels can transmit data, and at what rate. In practice, noiseless and noisy codes are combined to transmit data across a noisy channel with high reliability and efficiency.

In addition to transmitting data reliably, we may also wish to transmit our message securely. This leads to the study of cryptography. There are several possible aims that a cryptographic cipher might try to achieve, for example ensuring that a message was not read or tampered with, or 'signing' a message to authenticate that it originated from a particular sender. These concepts form the basis of modern internet security.

Contents

1. Modelling communication

To reason about communication, we use the following model. We have a *source* which knows a message, that uses an *encoder* to produce some *code words*. The code words are sent through a *channel*, but errors and noise may be introduced in this channel. The code words are received by a *decoder*, which performs some form of error detection and correction. The message is finally received by a *receiver*.

The source is often named *Alice*, and the receiver is named *Bob*. There may be an agent watching the channel called *Eve*, short for eavesdropper.

Examples of these ideas include the optical and electrical telegraph, SMS, postcodes, CDs and their error correction, compression algorithms such as `gzip`, and PINs.

Given a source and a channel, modelled probabilistically, the basic problem is to design an encoder and decoder to transmit messages *economically* (noiseless coding, compression) and *reliably* (noisy coding).

An example of noiseless coding is Morse code, where every letter is assigned a unique sequence of dots and dashes, where more common letters are assigned shorter strings. Noiseless coding is adapted to the source.

Here is an example of noisy coding. Each book has an ISBN $a_1a_2 \dots a_9a_{10}$ where the $a_1, \dots, a_9$ are digits in $\{0, \dots, 9\}$, and $a_{10} \in \{0, \dots, 9, X\}$ such that $11 \mid \sum_{j=1}^{10} ja_j$. This coding system detects the common human errors of writing an incorrect digit and transposing two adjacent digits. Noisy coding is adapted to the channel, which in this case is the human reading the number and typing it into a computer.

Definition. A *communication channel* accepts a string of symbols from a finite alphabet $\mathcal{A} = \{a_1, \dots, a_r\}$ and outputs a string of symbols from another finite alphabet $\mathcal{B} = \{b_1, \dots, b_s\}$. It is modelled by the probabilities $\mathbb{P}(y_1 \dots y_n \text{ received} \mid x_1 \dots x_n \text{ sent})$.

Definition. A *discrete memoryless channel* is a channel where $p_{ij} = \mathbb{P}\left(b_j \text{ received} \mid a_i \text{ sent}\right)$ are the same for each channel use, and independent of all past and future uses of the channel. Its *channel matrix* is the $r \times s$ stochastic matrix $P = (p_{ij})$.

Example. The *binary symmetric channel* with error probability $p \in [0,1]$ is a discrete memoryless channel with input and output alphabets $\{0,1\}$, where the channel matrix is

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

Here, a symbol is transmitted correctly with probability $1-p$. Usually, we assume $p < \frac{1}{2}$.

Example. The *binary erasure channel* has $\mathcal{A} = \{0,1\}$ and $\mathcal{B} = \{0,1,\star\}$. The channel matrix is

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

p can be interpreted as the probability that the symbol received is unreadable. If $\star$ is received, we say that we have received a *splurge error*.

Definition. We model n uses of a channel by the *nth extension*, with input alphabet $\mathcal{A}^n$ and output alphabet $\mathcal{B}^n$. A *code* C of length n is a function $\mathcal{M} \to \mathcal{A}^n$, where $\mathcal{M}$ is a set of messages. Implicitly, we also have a decoding rule $\mathcal{B}^n \to \mathcal{M}$.

- The *size* of this code is $m = |\mathcal{M}|$.
- The *information rate* of the code is $\rho(C) = \frac{1}{n} \log_2 m$.
- The *error rate* of the code is $\hat{e}(C) = \max_{x \in \mathcal{M}} \mathbb{P}\left(\text{error} \mid x \text{ sent}\right)$.

Definition. A channel can *transmit reliably at rate R* if there is a sequence of codes $(C_n)_{n=1}^{\infty}$ with each C_n a code of length n such that $\lim_{n\to\infty} \rho(C_n) = R$ and $\lim_{n\to\infty} \hat{e}(C_n) = 0$. The *capacity* of a channel is the supremum of all reliable transmission rates.

It is a nontrivial fact that the capacity of the binary symmetric channel with $p < \frac{1}{2}$ is nonzero. This is one of Shannon's theorems, proven later.

2. Noiseless coding

2.1. Prefix-free codes

Let $\mathcal{A}$ be a finite alphabet. We write $\mathcal{A}^\star$ for the set of strings of elements of $\mathcal{A}$, defined by $\mathcal{A}^\star = \bigcup_{n \geq 0} A^n$. The *concatenation* of two strings $x = x_1 \dots x_r$ and $y = y_1 \dots y_s$ is the string $xy = x_1 \dots x_r y_1 \dots y_s$.

Definition. Let $\mathcal{A}, \mathcal{B}$ be alphabets. A *code* is a function $c \colon \mathcal{A} \to \mathcal{B}^\star$. The *codewords* of c are the elements of $\operatorname{Im} c$.

Example (Greek fire code)**.** Let $\mathcal{A} = \{\alpha, \beta, \dots, \omega\}$, and $B = \{1, 2, 3, 4, 5\}$. We map $c(\alpha) = 11, c(\beta) = 12, \dots, c(\psi) = 53, c(\omega) = 54$. xy means to hold up x torches and another y torches nearby. This code was described by the historian Polybius.

Example. Let $\mathcal{A}$ be a set of words in some dictionary. Let $\mathcal{B}$ be the letters of English $\{A, \dots, Z, \text{␣}\}$ The code is to spell the word and follow with a space.

The general idea is to send a message $x_1, \dots, x_n \in \mathcal{A}^\star$ as $c(x_1) \dots c(x_n) \in \mathcal{B}^\star$. So c extends to a function $c^\star : \mathcal{A}^\star \to \mathcal{B}^\star$.

Definition. A code c is *decipherable* (or *uniquely decodable*) if $c^\star$ is injective.

If c is decipherable, each string in $\mathcal{B}^\star$ corresponds to at most one message. It does not suffice to require that c be injective. Consider $\mathcal{A} = \{1, 2, 3, 4\}, \mathcal{B} = \{0, 1\}$, and let $c(1) = 0, c(2) = 1, c(3) = 00, c(4) = 01$. Then $c^\star(114) = 0001 = c^\star(312)$.

Typically we define $m = |\mathcal{A}|$ and $a = |\mathcal{B}|$. We say c is an a-ary code of size m. A 2-ary code is a binary code, and a 3-ary code is a ternary code. We aim to construct decipherable codes with short word lengths. Assuming that c is injective, the following codes are always decipherable.

(i) a *block code*, where all codewords have the same length, such as in the Greek fire code;

(ii) a *comma code*, which reserves a letter from $\mathcal{B}$ to signal the end of a word;

(iii) a *prefix-free code*, a code in which no codeword is a prefix of another codeword.

Block codes and comma codes are examples of prefix-free codes. Such codes require no lookahead to determine if we have reached the end of a word, so such codes are sometimes called *instantaneous* codes. One can easily find decipherable codes that are not prefix-free.

2.2. Kraft's inequality

Definition. Let $\mathcal{A}$ be an alphabet of size m, and $\mathcal{B}$ be an alphabet of size a. Let $c \colon \mathcal{A} \to \mathcal{B}^\star$ be a code with codewords are of length $\ell_1, \dots, \ell_m$. Then, *Kraft's inequality* is

$$\sum_{i=1}^m a^{-\ell_i} \leq 1$$

Theorem. A prefix-free code (with given codeword lengths) exists if and only if Kraft's inequality holds.

Proof. Let us rewrite Kraft's inequality as $\sum_{\ell=1}^{s} n_\ell a^{-\ell} \leq 1$, where n_ℓ is the number of codewords of length ℓ, and s is the length of the longest codeword. Suppose $c : \mathcal{A} \to \mathcal{B}^\star$ is prefix-free. Then,

$$n_1 a^{s-1} + n_2 a^{s-2} + \dots + n_{s-1} a + n_s \leq a^s$$

since the left hand side counts the number of strings of length s in $\mathcal{B}$ with some codeword of c as a prefix, and the right hand side counts the total number of strings of length s. Dividing by a^s gives the desired result.

Now, suppose that $\sum_{\ell=1}^{s} n_\ell a^{-\ell} \leq 1$. We aim to construct a prefix-free code c with n_ℓ codewords of length ℓ for all $\ell \leq s$. Proceed by induction on s. The case $s = 1$ is clear; in this case, the inequality gives $n_1 \leq a$. By the inductive hypothesis, we have constructed a prefix-free code $\hat{c}$ with n_ℓ codewords of length ℓ for all $\ell < s$. The inequality gives $n_1 a^{s-1} + \dots + n_{s-1} a + n_s \leq a^s$. The first $s - 1$ terms on the left hand side gives the number of strings of length s with some codeword of $\hat{c}$ as a prefix. So we are free to add n_s additional codewords of length s to $\hat{c}$ to form c without exhausting our supply of a^s total strings of length s. □

Remark. The proof of existence of such a code is constructive; one can choose codewords in order of increasing length, ensuring that we do not introduce prefixes at each stage.

2.3. McMillan's inequality

Theorem. Any decipherable code satisfies Kraft's inequality.

Proof. Let $c : \mathcal{A} \to \mathcal{B}^\star$ be decipherable with word lengths $\ell_1, \dots, \ell_m$. Let $s = \max_{i \leq m} \ell_i$. For $R \in \mathbb{N}$, we have

$$\left(\sum_{i=1}^{m} a^{-\ell_i} \right)^R = \sum_{\ell=1}^{Rs} b_\ell a^{-\ell}$$

where b_ℓ is the number of ways of choosing R codewords of total length ℓ. Since c is decipherable, any string of length ℓ formed from codewords must correspond to exactly one sequence of codewords. Hence, $b_\ell \leq |\mathcal{B}^\ell| = a^\ell$. The inequality therefore gives

$$\left(\sum_{i=1}^{m} a^{-\ell_i} \right)^R \leq Rs \implies \sum_{i=1}^{m} a^{-\ell_i} \leq (Rs)^{\frac{1}{R}}$$

As $R \to \infty$, the right hand side converges to 1, giving Kraft's inequality as required. □

Corollary. A decipherable code with prescribed word lengths exists if and only if a prefix-free code with the same word lengths exists.

We can therefore restrict our attention to prefix-free codes.

2.4. Entropy

Entropy is a measure of 'randomness' or 'uncertainty' in an input message. Suppose that we have a random variable X taking a finite number of values $x_1, \dots, x_n$ with probability $p_1, \dots, p_n$. Then, the entropy of this random variable is the expected number of fair coin tosses required to determine X.

Example. Suppose $p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$. Identifying $\{x_1, x_2, x_3, x_4\} = \{00, 01, 10, 11\}$, we would expect $H(X) = 2$.

Example. Suppose $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{4}$, and $p_3 = p_4 = \frac{1}{8}$. Identifying $\{x_1, x_2, x_3, x_4\} = \{0, 10, 110, 111\}$, we obtain $H(X) = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 = \frac{7}{4}$.

In a sense, the first example is 'more random' than the second, as its entropy is higher.

Definition. The *entropy* of a random variable X taking a finite number of values $x_1, \dots, x_n$ with probabilities $p_1, \dots, p_n$ is defined to be

$$H(X) = H(p_1, \dots, p_n) = -\sum_{i=1}^{n} p_i \log p_i = -\mathbb{E}\left[\log p_i\right]$$

where the logarithm is taken with base 2.

Note that $H(X) \geq 0$, and equality holds exactly when X is constant with probability 1. It is measured in *bits*, binary digits. By convention, we write $0 \log 0 = 0$ (note that $x \log x \to 0$ as $x \to 0$).

Example. For a biased coin with probability p of a head, we write $H(p, 1 - p) = H(p)$. We find

$$H(p) = -p \log p - (1 - p) \log(1 - p); \quad H'(p) = \log \frac{1 - p}{p}$$

This graph is concave, taking a maximum value of 1 when $p = \frac{1}{2}$. If $p = 0, 1$ then $H(p) = 0$.

2.5. Gibbs' inequality

Proposition. Let $(p_1, \dots, p_n), (q_1, \dots, q_n)$ be discrete probability distributions. Then,

$$-\sum p_i \log p_i \leq -\sum p_i \log q_i$$

with equality if and only if $p_i = q_i$.

The right hand side is sometimes called the *cross entropy*, or *mixed entropy*.

Proof. Since $\log x = \frac{\ln x}{\ln 2}$, we may replace the inequality with

$$-\sum p_i \ln p_i \leq -\sum p_i \ln q_i$$

Define $I = \{i \mid p_i \neq 0\}$. Now, $\ln x \leq x - 1$ for all $x > 0$, with equality if and only if $x = 1$. Hence, $\ln \frac{q_i}{p_i} \leq \frac{q_i}{p_i} - 1$ for all $i \in I$. Then,

$$\sum_{i\in I} p_i \ln \frac{q_i}{p_i} \leq \sum_{i\in I} q_i - \sum_{i\in I} p_i$$

As the p_i form a probability distribution, $\sum_{i\in I} p_i = 1$ and $\sum_{i\in I} q_i \leq 1$, so the right hand side is at most 0. Therefore,

$$-\sum_{i=1}^{n} p_i \ln p_i = -\sum_{i\in I} p_i \ln p_i \leq -\sum_{i\in I} p_i \ln q_i \leq -\sum_{i=1}^{n} p_i \ln q_i$$

If equality holds, we must have $\sum_{i\in I} q_i = 1$ and $\frac{q_i}{p_i} = 1$ for all $i \in I$, giving that $p_i = q_i$ for all i. □

Corollary. $H(p_1, \dots, p_n) \leq \log n$, with equality if and only if $p_1 = \dots = p_n$.

2.6. Optimal codes

Let $\mathcal{A} = \{\mu_1, \dots, \mu_m\}$ be an alphabet of $m \geq 2$ messages, and let $\mathcal{B}$ be an alphabet of length $a \geq 2$. Let X be a random variable taking values in A with probabilities $p_1, \dots, p_m$.

Definition. A code $c \colon \mathcal{A} \to \mathcal{B}^\star$ is called *optimal* if it has the smallest possible expected word length $\sum p_i \ell_i = \mathbb{E}[S]$ among all decipherable codes.

Theorem (Shannon's noiseless coding theorem). The expected word length $\mathbb{E}[S]$ of a decipherable code satisfies

$$\overbrace{\frac{H(X)}{\log a} \leq \mathbb{E}[S]}^{\text{for decipherable codes}} \underbrace{< \frac{H(X)}{\log a} + 1}_{\text{for optimal codes}}$$

Moreover, the left hand inequality is an equality if and only if $p_i = a^{-\ell_i}$ with $\sum a^{-\ell_i} = 1$ for some integers $\ell_1, \dots, \ell_m$.

Proof. First, we consider the lower bound. Let $c \colon \mathcal{A} \to \mathcal{B}^\star$ be a decipherable code with word lengths $\ell_1, \dots, \ell_m$. Let $q_i = \frac{a^{-\ell_i}}{D}$ where $D = \sum a^{-\ell_i}$, so $\sum q_i = 1$. By Gibbs' inequality,

$$H(X) \leq -\sum p_i \log q_i = -\sum p_i(-\ell_i \log a - \log D) = \log D + \log a \sum p_i \ell_i$$

By McMillan's inequality, $D \leq 1$ so $\log D \leq 0$. Hence, $H(X) \leq \log a \sum p_i \ell_i = \log a \mathbb{E}[S]$ as required. Equality holds exactly when $D = 1$ and $p_i = q_i = \frac{a^{-\ell_i}}{D} = a^{-\ell_i}$ for some integers $\ell_1, \dots, \ell_m$.

Now, consider the upper bound. We construct a code called the *Shannon–Fano code*. Let $\ell_i = \lceil -\log_a p_i \rceil$, so $-\log_a p_i \leq \ell_i < -\log_a p_i + 1$. Therefore, $\log_a p_i \geq -\ell_i$, so $p_i \geq a^{-\ell_i}$.

Thus, Kraft's inequality $\sum a^{-\ell_i} \leq 1$ is satisfied, so there exists a prefix-free code c with these word lengths $\ell_1, \dots, \ell_m$. c has expected word length

$$\mathbb{E}\left[S\right] = \sum p_i \ell_i < \sum p_i(-\log p_i + 1) = \frac{H(X)}{\log a} + 1$$

as required. □

Example (Shannon–Fano coding). For probabilities $p_1, \dots, p_m$, we set $\ell_i = \lceil -\log_a p_i \rceil$. Construct a prefix-free code with these word lengths by choosing codewords in order of size, with smallest codewords being selected first to ensure that the prefix-free property holds. By Kraft's inequality, this process can always be completed.

Example. Let $a = 2, m = 5$, and define

i	p_i	$\lceil -\log_2 p_i \rceil$	
1	0.4	2	00
2	0.2	3	010
3	0.2	3	011
4	0.1	4	1000
5	0.1	4	1001

Here, $\mathbb{E}\left[S\right] = \sum p_i \ell_i = 2.8$, and $H(X) = \frac{H(X)}{\log 2} \approx 2.12$. Clearly, this is not optimal; one could take $c(4) = 100, c(5) = 101$ to reduce the expected word length.

2.7. Huffman coding

Let $\mathcal{A} = \{\mu_1, \dots, \mu_m\}$ and $p_i = \mathbb{P}\left(X = \mu_i\right)$. We assume $a = 2$ and $\mathcal{B} = \{0, 1\}$ for simplicity. Without loss of generality, we can assume $p_1 \geq p_2 \geq \dots \geq p_m$. We construct an optimal code inductively.

If $m = 2$, we take codewords 0 and 1. If $m > 2$, first we take the Huffman code for messages $\mu_1, \dots, \mu_{m-2}, \nu$ with probabilities $p_1, p_2, \dots, p_{m-2}, p_{m-1} + p_m$. Then, we append 0 and 1 to the codeword for ν to obtain the new codewords for μ_{m-1}, μ_m.

Remark. By construction, Huffman codes are prefix-free. In general, Huffman codes are not unique; we require a choice if $p_i = p_j$.

Example. Consider the example Let $a = 2, m = 5$, and consider as before

i	p_i
1	0.4
2	0.2
3	0.2
4	0.1
5	0.1

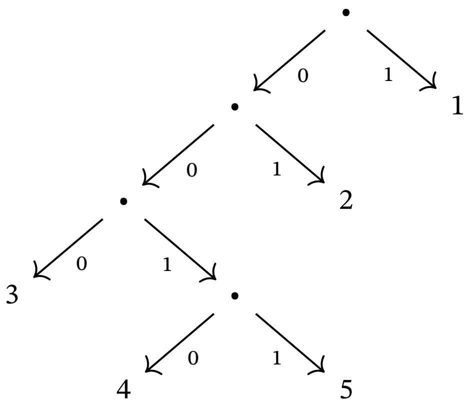
Merging 4 and 5, as they have the lowest probabilities,

i	p_i
1	0.4
2	0.2
3	0.2
45	0.2

Continuing, we obtain

i	p_i
(3(45))2	0.6
1	0.4

giving codewords



This gives $\mathbb{E}[S] = 2.2$, better than the Shannon–Fano code found above.

Lemma. Let $\mu_1, \dots, \mu_m$ be messages in $\mathcal{A}$ with probabilities $p_1, \dots, p_m$. Let c be an optimal prefix-free code for c with word lengths $\ell_1, \dots, \ell_m$. Then,

(i) if $p_i > p_j$, $\ell_i \leq \ell_j$; and

(ii) among all codewords of maximal length, there exist two which differ only in the last digit.

Proof. If this were not true, one could modify c by

(i) swapping the ith and jth codewords; or

(ii) deleting the last letter of each codeword of maximal length

which yields a prefix-free code with strictly smaller expected word length. □

Theorem. Huffman codes are optimal.

Proof. The proof is by induction on m. If $m = 2$, then the codewords are 0 and 1, which is clearly optimal. Assume $m > 2$, and let c_m be the Huffman code for X_m which takes values $\mu_1, \dots, \mu_m$ with probabilities $p_1 \geq \dots \geq p_m$. c_m is constructed from a Huffman code c_{m-1}

with random variable X_{m-1} taking values $\mu_1, \dots, \mu_{n-2}, \nu$ with probabilities $p_1, \dots, p_{m-2}, p_{m-1} + p_m$. The code c_{m-1} is optimal by the inductive hypothesis. The expected word length $\mathbb{E}[S_m]$ is given by

$$\mathbb{E}[S_m] = \mathbb{E}[S_{m-1}] + p_{m-1} + p_m$$

Let c'_m be an optimal code for X_m, which without loss of generality can be chosen to be prefix-free. Without loss of generality, the last two codewords of c'_m can be chosen to have the largest possible length and differ only in the final position, by the previous lemma. Then, $c'_m(\mu_{m-1}) = y0$ and $c'_m(\mu_m) = y1$ for some $y \in \{0,1\}^*$. Let c'_{m-1} be the prefix-free code for X_{m-1} given by

$$c'_{m-1}(\mu_i) = \begin{cases} c'_m(\mu_i) & i \leq m-2 \\ y & i = m-1, m \end{cases}$$

The expected word length satisfies

$$\mathbb{E}[S'_m] = \mathbb{E}[S'_{m-1}] + p_{m-1} + p_m$$

By the inductive hypothesis, c_{m-1} is optimal, so $\mathbb{E}[S_{m-1}] \leq \mathbb{E}[S'_{m-1}]$. Combining the equations,

$$\mathbb{E}[S_m] \leq \mathbb{E}[S'_m]$$

So c_m is optimal as required. □

Remark. Not all optimal codes are Huffman codes. However, we have proven that, given a prefix-free optimal code with prescribed word lengths, there is a Huffman code with these word lengths.

2.8. Joint entropy

Let X, Y be random variables with values in $\mathcal{A}, \mathcal{B}$. Then, the pair (X, Y) is also a random variable, taking values in $\mathcal{A} \times \mathcal{B}$. This has entropy $H(X, Y)$, called the *joint entropy* for X and Y.

$$H(X,Y) = -\sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} \mathbb{P}(X = x, Y = y) \log \mathbb{P}(X = x, Y = y)$$

This construction generalises to finite tuples of random variables.

Lemma. Let X, Y be random variables taking values in $\mathcal{A}, \mathcal{B}$. Then $H(X, Y) \leq H(X) + H(Y)$, with equality if and only if X and Y are independent.

Proof. Let $\mathcal{A} = \{x_1, \dots, x_m\}$ and $\mathcal{B} = \{y_1, \dots, y_n\}$. Let $p_{ij} = \mathbb{P}(X = x_i, Y = y_j)$, $p_i =$

$\mathbb{P}(X = x_i)$, and $q_j = \mathbb{P}\left(Y = y_j\right)$. By Gibbs' inequality applied to $\left\{p_{ij}\right\}$ and $\left\{p_i q_j\right\}$,

$$\begin{aligned} H(X, Y) = -\sum p_{ij} \log p_{ij} &\leq -\sum p_{ij} \log\left(p_i q_j\right) \\ &= -\sum_i \left(\sum_j p_{ij}\right) \log p_i - \sum_j \left(\sum_i p_{ij}\right) \log q_j \\ &= -\sum_i p_i \log p_i - \sum_j q_j \log q_j \\ &= H(X) + H(Y) \end{aligned}$$

Equality holds if and only if $p_{ij} = p_i q_j$ for all i, j, or equivalently, if X, Y are independent. □

3. Noisy channels

3.1. Decoding rules

Definition. A *binary* $[n, m]$*-code* is a subset C of $\{0,1\}^n$ of size $m = |C|$. We say n is the *length* of the code, and elements of C are called *codewords*.

We use an $[n, m]$-code to send one of m messages through a channel using n bits. For instance, if the channel is a binary symmetric channel, we use the channel n times. Note that $1 \leq m \leq 2^n$, so the information rate $\rho(C) = \frac{1}{n}\log m$ satisfies $0 \leq \rho(C) \leq 1$. If $m = 1$, $\rho(C) = 0$, and if $C = \{0,1\}^n$, $\rho(C) = 1$.

Definition. Let $x, y \in \{0,1\}^n$. The *Hamming distance* between x and y is

$$d(x, y) = |\{i \mid x_i \neq y_i\}|$$

In this section, we consider only the binary symmetric channel with probability p.

Definition. Let C be a binary $[n, m]$-code.

- The *ideal observer* decoding rule decodes $x \in \{0,1\}^n$ as the $c \in C$ maximising the probability that c was sent given that x was received;
- The *maximum likelihood* decoding rule decodes $x \in \{0,1\}^n$ as the $c \in C$ maximising the probability that x was received given that c was sent;
- The *minimum distance* decoding rule decodes $x \in \{0,1\}^n$ as the $c \in C$ minimising the Hamming distance $d(x, c)$.

Lemma. Let C be a binary $[n, m]$-code.

(i) If all messages are equally likely, the ideal observer and maximum likelihood decoding rules agree.

(ii) If $p < \frac{1}{2}$, then the maximum likelihood and minimum distance decoding rules agree.

Note that the hypothesis in part (i) is reasonable if we first encode a message using noiseless coding. The hypothesis in part (ii) is reasonable, since a channel with $p = \frac{1}{2}$ can carry no information, and a channel with $p > \frac{1}{2}$ can be used as a channel with probability $1 - p$ by inverting its outputs. Channels with $p = 0$ are called *lossless channels*, and channels with $p = \frac{1}{2}$ are called *useless channels*.

Proof. Part (i). By Bayes' rule,

$$\mathbb{P}\left(c \text{ sent} \mid x \text{ received}\right) = \frac{\mathbb{P}\left(c \text{ sent}, x \text{ received}\right)}{x \text{ received}} = \frac{\mathbb{P}\left(c \text{ sent}\right)}{\mathbb{P}\left(x \text{ received}\right)}\mathbb{P}\left(x \text{ received} \mid c \text{ sent}\right)$$

By hypothesis, $\mathbb{P}(c \text{ sent})$ is independent of c. Hence, for some fixed received message x, maximising $\mathbb{P}(c \text{ sent} \mid x \text{ received})$ is the same as maximising $\mathbb{P}(x \text{ received} \mid c \text{ sent})$.

Part (ii). Let $r = d(x,c)$. Then,

$$\mathbb{P}\,(x \text{ received} \mid c \text{ sent}) = p^r(1-p)^{n-r} = (1-p)^n\left(\frac{p}{1-p}\right)^r$$

As $p < \frac{1}{2}$, $\frac{p}{1-p} < 1$. Hence, maximising $\mathbb{P}\,(x \text{ received} \mid c \text{ sent})$ is equivalent to minimising $r = d(x,c)$. □

We can therefore choose to use minimum distance decoding from this point.

Example. Suppose codewords 000, 111 are sent with probabilities $\alpha = \frac{9}{10}$ and $1 - \alpha = \frac{1}{10}$, through a binary symmetric channel with error probability $p = \frac{1}{4}$. Suppose that we receive 110. Clearly, an error has been introduced.

$$\begin{aligned}\mathbb{P}\,(000 \text{ sent} \mid 110 \text{ received}) &= \frac{\alpha p^2(1-p)}{\alpha p^2(1-p) + (1-\alpha)p(1-p)^2} = \frac{3}{4}\\ \mathbb{P}\,(111 \text{ sent} \mid 110 \text{ received}) &= \frac{1}{4}\end{aligned}$$

The ideal observer therefore decodes 110 as 000. The maximum likelihood or minimum distance decoding rules decode 110 as 111.

Remark. Minimum distance decoding may be expensive in terms of time and storage if $|C|$ is large, since the distance to all codewords must be calculated *a priori*. One must also specify a convention in case of a tie between the probabilities or distances, for instance, using a random choice, or requesting a retransmission.

3.2. Error detection and correction

The aim when constructing codes for noisy channels is to detect errors, and if possible, to correct them.

Definition. A binary $[n,m]$-code C is

- *d-error detecting* if, when changing up to d digits in each codeword, we can never produce another codeword;
- *e-error correcting* if, knowing that $x \in \{0,1\}^n$ differs from a codeword in at most e positions, we can deduce the codeword.

Example. A *repetition code* of length n has codewords $0^n, 1^n$. This is an $[n,2]$-code. It is $(n-1)$-error detecting, and $\left\lfloor \frac{n-1}{2} \right\rfloor$-error correcting. Its information rate is $\frac{1}{n}$.

Example. A *simple parity check code* or *paper tape code* of length n identifies the set $\{0,1\}$ with the field $\mathbb{F}_2$ of two elements, and defines $C = \{(x_1, \dots, x_n) \in \mathbb{F}_2^n \mid \sum x_i = 0\}$. This is an $[n, 2^{n-1}]$-code. This is 1-error detecting and 0-error correcting, but has information rate $\frac{n-1}{n}$.

Example. Hamming's original code is a 1-error correcting binary [7, 16]-code, defined on a subset of $\mathbb{F}_2^7$ by

$$C = \left\{c \in \mathbb{F}_2^7 \mid c_1 + c_3 + c_5 + c_7 = 0; c_2 + c_3 + c_6 + c_7 = 0; c_4 + c_5 + c_6 + c_7 = 0\right\}$$

The bits c_3, c_5, c_6, c_7 are chosen arbitrarily, and c_1, c_2, c_4 are check digits, giving a size of $2^4 = 16$. Suppose that we receive $x \in \mathbb{F}_2^7$. We form the *syndrome* $z = z_x = (z_1, z_2, z_4) \in \mathbb{F}_2^3$ where

$$z_1 = x_1 + x_3 + x_5 + x_7; \quad z_2 = x_2 + x_3 + x_6 + x_7; \quad z_4 = x_4 + x_5 + x_6 + x_7$$

By definition of C, if $x \in C$ then $z = (0, 0, 0)$. If $d(x, c) = 1$ for some $c \in C$, then the place where x and c differ is given by $z_1 + 2z_2 + 4z_4$ (not modulo 2). Indeed, if $x = c + e_i$ where e_i is the zero vector with a one in the ith position, $z_x = z_{e_i}$, and one can check that this holds for each $1 \leq i \leq 7$. Therefore, Hamming's original code is 1-error correcting.

Lemma. The Hamming distance is a metric on $\mathbb{F}_2^n$.

Proof. Clearly, $d(x, y) \geq 0$ and equality holds if and only if $x = y$, and $d(x, y) = d(y, x)$. Let $x, y, z \in \mathbb{F}_2^n$. Then,

$$\{i \mid x_i \neq z_i\} \subseteq \{i \mid x_i \neq y_i\} \cup \{i \mid y_i \neq z_i\}$$

Hence $d(x, z) \leq d(x, y) + d(y, z)$. □

Remark. We could write $d(x, y)$ as $\sum d_1(x_i, y_i)$ where d_1 is the discrete metric on $\mathbb{F}_2$.

3.3. Minimum distance

Definition. The *minimum distance* of a code is the minimum value of $d(c_1, c_2)$ for codewords $c_1 \neq c_2$.

Lemma. Let C be a code with minimum distance $d > 0$. Then,

(i) C is $(d-1)$-error detecting, but cannot detect all sets of d errors;

(ii) C is $\left\lfloor \frac{d-1}{2} \right\rfloor$-error correcting, but cannot correct all sets of $\left\lfloor \frac{d-1}{2} \right\rfloor + 1$ errors.

Proof. Part (i). If $x \in \mathbb{F}_2^n$ and c is a codeword with $1 \leq d(x, c) \leq d - 1$. Then $x \notin C$, so $d - 1$ errors are detected. Suppose c_1, c_2 are codewords with $d(c_1, c_2) = d$. Then c_1 can be corrupted into c_2 with only d errors, and this is undetectable.

Part (ii). Let $e = \left\lfloor \frac{d-1}{2} \right\rfloor$. By definition, $e \leq \frac{d-1}{2} < e + 1$, so $2e < d \leq 2(e + 1)$. Let $x \in \mathbb{F}_2^n$. If $c_1 \in C$ with $d(x, c_1) \leq e$, we want to show that $d(x, c_2) > e$ for all $c_2 \neq c_1$. By the triangle inequality, $d(x, c_2) \geq d(c_1, c_2) - d(x, c_1) \geq d - e > e$ as required. Hence, C is e-error correcting.

Let $c_1, c_2 \in C$ with $d(c_1, c_2) = d$. Let $x \in \mathbb{F}_2^n$ differ from c_1 in precisely $e + 1$ places that c_1 and c_2 differ. Then $d(x, c_1) = e + 1$, and $d(x, c_2) = d - (e + 1) \leq e + 1$. Hence, C cannot correct all sets of $e + 1$ errors. □

Definition. An $[n, m]$-code with minimum distance d is called an $[n, m, d]$-code.

Note that $m \leq 2^n$ with equality if and only if $C = \mathbb{F}_2^n$. Similarly, $d \leq n$, with equality in the case of the repetition code.

Example. The repetition code of length n is an $[n, 2, n]$-code. The simple parity check code of length n is an $[n, 2^{n-1}, 2]$-code. The trivial code on n bits is an $[n, 2^n, 1]$-code. Hamming's original code is 1-error correcting, so has minimum distance at least 3. The minimum distance can easily be shown to be exactly 3 as $0000000, 1110000$ are codewords, so it is a $[7, 16, 3]$-code.

3.4. Covering estimates

Definition. Let $x \in \mathbb{F}_2^n$ and $r \geq 0$. Then, we denote the *closed Hamming ball* with centre x and radius r by $B(x, r)$. We write $V(n, r) = |B(x, r)| = \sum_{i=0}^{r} \binom{n}{i}$ for the *volume* of this ball.

Lemma (Hamming's bound; sphere packing bound)**.** An e-error correcting code C of length n has

$$|C| \leq \frac{2^n}{V(n, e)}$$

Proof. C is e-error correcting, so $B(c_1, e) \cap B(c_2, e)$ is empty for all codewords $c_1 \neq c_2$. Hence,

$$\sum_{c \in C} |B(c, e)| \leq |\mathbb{F}_2^n| \implies |C| V(n, e) \leq 2^n$$

as required. □

Definition. An e-error correcting code C of length n such that $|C| = \frac{2^n}{V(n,e)}$ is called *perfect*.

Remark. Equivalently, a code is perfect if for all $x \in \mathbb{F}_2^n$, there exists a unique $c \in C$ such that $d(x, c) \leq e$. Alternatively, $\mathbb{F}_2^n$ is a union of disjoint balls $B(c, e)$ for all $c \in C$, or that any collection of $e + 1$ will cause the message to be decoded incorrectly.

Example. Consider Hamming's $[7, 16, 3]$-code. This is 1-error correcting, and

$$\frac{2^n}{V(n, e)} = \frac{2^7}{V(7, 1)} = \frac{2^7}{1 + 7} = 2^4 = |C|$$

So Hamming's original code is perfect.

Example. The binary repetition code of length n is perfect if and only if n is odd.

Remark. If $\frac{2^n}{V(n,e)}$ is not an integer, there does not exist a perfect e-error correcting code of length n. The converse is false; the case $n = 90, e = 2$ is discussed on the second example sheet.

Definition. $A(n, d)$ is the largest possible size m of an $[n, m, d]$-code.

The values of the $A(n, d)$ are unknown in general.

Example. $A(n, 1) = 2^n$, considering the trivial code. $A(n, 2) = 2^{n-1}$, maximised at the simple parity check code. $A(n, n) = 2$, maximised at the repetition code.

Lemma. $A(n, d + 1) \leq A(n, d)$.

Proof. Let $m = A(n, d + 1)$, and let C be an $[n, m, d + 1]$-code. Let $c_1, c_2 \in C$ be distinct codewords such that $d(c_1, c_2) = d + 1$. Let c_1' differ from c_1 in exactly one of the places where c_1 and c_2 differ. Then $d(c_1', c_2) = d$. If $c \in C$ is any codeword not equal to c_1, then $d(c, c_1) \leq d(c, c_1') + d(c_1', c_1)$ hence $d + 1 \leq d(c, c_1') + 1$, so the code given by $C \cup \{c_1'\} \setminus \{c_1\}$ has minimum distance d, but has length n and size m. This is therefore an $[n, m, d]$-code as required. □

Corollary. Equivalently, $A(n, d) = \max\{m \mid \exists[n, m, d']\text{-code, for some } d' \geq d\}$.

Theorem.

$$\frac{2^n}{V(n, d-1)} \leq A(n, d) \leq \frac{2^n}{V\left(n, \left\lfloor \frac{d-1}{2} \right\rfloor\right)}$$

The upper bound is Hamming's bound; the lower bound is known as the GSV (Gilbert–Shannon–Varshamov) bound. The upper bound can be thought of as a sphere packing bound, and the lower bound is a sphere covering bound.

Proof. We prove the lower bound. Let $m = A(n, d)$, and let C be an $[n, m, d]$-code. Then, there exists no $x \in \mathbb{F}_2^n$ with $d(x, c) \geq d$ for all codewords. Indeed, if such an x exists, we could consider the code $C \cup \{x\}$, which would be an $[n, m + 1, d]$-code, contradicting maximality of m. Then,

$$\mathbb{F}_2^n \subseteq \bigcup_{c \in C} B(c, d-1) \implies 2^n \leq \sum_{c \in C} |B(c, d-1)| = mV(n, d-1)$$

as required. □

Example. Let $n = 10, d = 3$. Then $V(n, 1) = 11$ and $V(n, 2) = 56$, so the GSV bound is $\frac{2^{10}}{56} \leq A(10, 3) \leq \frac{2^{10}}{11}$. Hence, $19 \leq A(10, 3) \leq 93$. It was known that the lower bound could be improved to 72. We now know that the true value of $A(10, 3)$ is exactly 72. In this case, the GSV bound was not a sharp inequality.

3.5. Asymptotics

We study the information rate $\frac{\log A(n, \lfloor n\delta \rfloor)}{n}$ as $n \to \infty$ to see how large the information rate can be for a fixed error rate.

Proposition. Let $0 < \delta < \frac{1}{2}$. Then,

(i) $\log V(n, \lfloor n\delta \rfloor) \leq nH(\delta)$;

(ii) $\frac{1}{n} \log A(n, \lfloor n\delta \rfloor) \geq 1 - H(\delta)$;

where $H(\delta) = -\delta \log \delta - (1-\delta)\log(1-\delta)$.

Proof. (i) implies (ii). By the GSV bound, we find

$$A(n, \lfloor n\delta \rfloor) \geq \frac{2^n}{V(n, \lfloor n\delta \rfloor - 1)} \geq \frac{2^n}{V(n, \lfloor n\delta \rfloor)}$$

Taking logarithms,

$$\frac{1}{n} \log A(n, \lfloor n\delta \rfloor) \geq 1 - \frac{\log V(n, \lfloor n\delta \rfloor)}{n} \geq 1 - H(\delta)$$

Part (i). $H(\delta)$ is increasing for $\delta < \frac{1}{2}$. Therefore, without loss of generality, we may assume $n\delta$ is an integer. Now, as $\frac{\delta}{1-\delta} < 1$,

$$\begin{aligned} 1 &= (\delta + (1-\delta))^n \\ &= \sum_{i=0}^{n} \binom{n}{i} \delta^i (1-\delta)^{n-i} \\ &\geq \sum_{i=0}^{n\delta} \binom{n}{i} \delta^i (1-\delta)^{n-i} \\ &= (1-\delta)^n \sum_{i=0}^{n\delta} \binom{n}{i} \left(\frac{\delta}{1-\delta}\right)^i \\ &\geq (1-\delta)^n \sum_{i=0}^{n\delta} \binom{n}{i} \left(\frac{\delta}{1-\delta}\right)^{n\delta} \\ &= \delta^{n\delta}(1-\delta)^{n(1-\delta)} V(n, n\delta) \end{aligned}$$

Taking logarithms,

$$0 \geq n\delta \log \delta + n(1-\delta)\log(1-\delta) + \log V(n, n\delta)$$

as required. □

The constant $H(\delta)$ in the proposition is optimal.

Lemma. $\lim_{n\to\infty} \frac{\log V(n, \lfloor n\delta \rfloor)}{n} = H(\delta)$.

Proof. Exercise. Follows from Stirling's approximation to factorials. □

3.6. Constructing new codes from old

Let C be an $[n, m, d]$-code.

Example. The *parity check extension* is an $[n+1, m, d']$-code given by

$$C^+ = \left\{ \left(c_1, \dots, c_n, \sum_{i=1}^n c_i \right) \middle| (c_1, \dots, c_n) \in C \right\}$$

where d' is either d or $d+1$, depending on whether d is odd or even.

Example. Let $1 \leq i \leq n$. Then, deleting the ith digit from each codeword gives the *punctured code*

$$C^- = \{(c_1, \dots, c_{i-1}, c_{i+1}, \dots, c_n) \mid (c_1, \dots, c_n) \in C\}$$

If $d \geq 2$, this is an $[n-1, m, d']$-code where d' is either d or $d-1$.

Example. Let $1 \leq i \leq n$ and let $\alpha \in \mathbb{F}_2$. The *shortened code* is

$$C' = \{(c_1, \dots, c_{i-1}, c_{i+1}, \dots, c_n) \mid (c_1, \dots, c_{i-1}, \alpha, c_{i+1}, \dots, c_n) \in C\}$$

This is an $[n-1, m', d']$ with $d' \geq d$ and $m' \geq \frac{m}{2}$ for a suitable choice of α.

4. Information theory

4.1. Sources and information rate

Definition. A *source* is a sequence of random variables $X_1, X_2, \dots$ taking values in $\mathcal{A}$.

Example. The *Bernoulli* (or *memoryless*) source is a source where the X_i are independent and identically distributed according to a Bernoulli distribution.

Definition. A source $X_1, X_2, \dots$ is *reliably encodable* at rate r if there exist subsets $A_n \subseteq \mathcal{A}^n$ such that

(i) $\lim \frac{\log |A_n|}{n} = r$;

(ii) $\lim \mathbb{P}\left((X_1, \dots, X_n) \in A_n\right) = 1$.

Definition. The *information rate* H of a source is the infimum of all reliable encoding rates.

Example. $0 \leq H \leq \log |\mathcal{A}|$, with both bounds attainable. The proof is left as an exercise.

Shannon's first coding theorem computes the information rate of certain sources, including Bernoulli sources.

Recall from IA Probability that a probability space is a tuple $(\Omega, \mathcal{F}, \mathbb{P})$, and a discrete random variable is a function $X : \Omega \to \mathcal{A}$. The probability mass function is the function $p_X : \mathcal{A} \to [0,1]$ given by $p_X(x) = \mathbb{P}(X = x)$. We can consider the function $p(X) : \Omega \to [0,1]$ defined by the composition $p_X \circ X$, which assigns $p(X)(\omega) = \mathbb{P}(X = X(\omega))$; hence, $p(X)$ is also a random variable.

Similarly, given a source $X_1, X_2, \dots$ of random variables with values in $\mathcal{A}$, the probability mass function of any tuple $X^{(n)} = (X_1, \dots, X_n)$ is $p_{X^{(n)}}(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$. As $p_{X^{(n)}} : \mathcal{A}^n \to [0,1]$, and $X^{(n)} : \Omega \to \mathcal{A}^n$, we can consider $p(X^{(n)}) = p_{X^{(n)}} \circ X^{(n)}$ defined by $\omega \mapsto p_{X^{(n)}}(X^{(n)}(\omega))$.

Example. Let $\mathcal{A} = \{A, B, C\}$. Suppose

$$X^{(2)} = \begin{cases} AB & \text{with probability } 0.3 \\ AC & \text{with probability } 0.1 \\ BC & \text{with probability } 0.1 \\ BA & \text{with probability } 0.2 \\ CA & \text{with probability } 0.25 \\ CB & \text{with probability } 0.05 \end{cases}$$

Then, $p_{X^{(2)}}(AB) = 0.3$, and so on. Hence,

$$p(X^{(2)}) = \begin{cases} 0.3 & \text{with probability } 0.3 \\ 0.1 & \text{with probability } 0.2 \\ 0.2 & \text{with probability } 0.2 \\ 0.25 & \text{with probability } 0.25 \\ 0.05 & \text{with probability } 0.05 \end{cases}$$

We say that a source $X_1, X_2, \dots$ converges in probability to a random variable L if for all $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\left(|X_n - L| > \varepsilon\right) = 0$. We write $X_n \xrightarrow{\mathbb{P}} L$. The weak law of large numbers states that if $X_1, X_2, \dots$ is a sequence of independent identically distributed real-valued random variables with finite expectation $\mathbb{E}[X_1]$, then $\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{\mathbb{P}} \mathbb{E}[X]$.

Example. Let $X_1, X_2, \dots$ be a Bernoulli source. Then $p(X_1), p(X_2), \dots$ are independent and identically distributed random variables, and $p(X_1, \dots, X_n) = p(X_1) \dots p(X_n)$. Note that by the weak law of large numbers,

$$-\frac{1}{n}\log p(X_1, \dots, X_n) = -\frac{1}{n}\sum_{i=1}^n \log p(X_i) \xrightarrow{\mathbb{P}} \mathbb{E}\left[-\log p(X_1)\right] = H(X_1)$$

Lemma. The information rate of a Bernoulli source $X_1, X_2, \dots$ is at most the expected word length of an optimal code $c : \mathcal{A} \to \{0,1\}^\star$ for X_1.

Proof. Let $\ell_1, \ell_2, \dots$ be the codeword lengths when we encode $X_1, X_2, \dots$ using c. Let $\varepsilon > 0$. Let

$$A_n = \{x \in \mathcal{A}^n \mid c^\star(x) \text{ has length less than } n(\mathbb{E}[\ell_1] + \varepsilon)\}$$

Then,

$$\mathbb{P}\left((X_1, \dots, X_n) \in A_n\right) = \mathbb{P}\left(\sum_{i=1}^n \ell_i \leq n(\mathbb{E}[\ell_1] + \varepsilon)\right) = \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \ell_i - \mathbb{E}[\ell_i]\right| < \varepsilon\right) \to 1$$

Now, c is decipherable so $c^\star$ is injective. Hence, $|A_n| \leq 2^{n(\mathbb{E}[\ell_1]+\varepsilon)}$. Making A_n larger if necessary, we can assume $|A_n| = \lfloor 2^{n(\mathbb{E}[\ell_1]+\varepsilon)} \rfloor$. Taking logarithms, $\frac{\log|A_n|}{n} \to \mathbb{E}[\ell_1] + \varepsilon$. So $X_1, X_2, \dots$ is reliably encodable at rate $r = \mathbb{E}[\ell_1] + \varepsilon$ for all $\varepsilon > 0$. Hence the information rate is at most $\mathbb{E}[\ell_1]$. □

Corollary. A Bernoulli source has information rate less than $H(X_1) + 1$.

Proof. Combine the previous lemma with the noiseless coding theorem. □

Suppose we encode $X_1, X_2, \dots$ in blocks of size N. Let $Y_1 = (X_1, \dots, X_N)$, $Y_2 = (X_{N+1}, \dots, X_{2N})$ and so on, such that $Y_1, Y_2, \dots$ take values in $\mathcal{A}^N$. One can show that if the source $X_1, X_2, \dots$ has information rate H, then $Y_1, Y_2, \dots$ has information rate NH.

Proposition. The information rate H of a Bernoulli source is at most $H(X_1)$.

Proof. Apply the previous corollary to the Y_i to obtain

$$NH < H(Y_1) + 1 = H(X_1, \dots, X_N) + 1 = NH(X_1) + 1 \implies H < H(X_1) + \frac{1}{N}$$

as required. □

4.2. Asymptotic equipartition property

Definition. A source $X_1, X_2, \dots$ satisfies the *asymptotic equipartition property* if there exists a constant $H \geq 0$ such that

$$-\frac{1}{n} \log p(X_1, \dots, X_n) \xrightarrow{\mathbb{P}} H$$

Example. Suppose we toss a biased coin with probability p of obtaining a head. Let $X_1, X_2, \dots$ be the results of independent coin tosses. If we toss the coin N times, we expect pN heads and $(1-p)N$ tails. The probability of any particular sequence of pN heads and $(1-p)N$ tails is

$$p^{pN}(1-p)^{(1-p)N} = 2^{N(p \log p + (1-p)\log(1-p))} = 2^{-NH(X)}$$

Not every sequence of tosses is of this form, but there is only a small probability of 'atypical sequences'. With high probability, it is a 'typical sequence' which has a probability close to $2^{-NH(X)}$.

Lemma. The asymptotic equipartition property for a source $X_1, X_2, \dots$ is equivalent to the property that for all $\varepsilon > 0$, there exists $n \in \mathbb{N}$ such that for all $n \geq n_0$, there exists a 'typical set' $T_n \subseteq \mathcal{A}^n$ such that

(i) $\mathbb{P}((X_1, \dots, X_n) \in T_n) > 1 - \varepsilon$;

(ii) $2^{-n(H+\varepsilon)} \leq p(x_1, \dots, x_n) \leq 2^{-n(H-\varepsilon)}$ for all $(x_1, \dots, x_n) \in T_n$.

Proof sketch. First, we show that the asymptotic equipartition property implies the alternative definition. We define

$$T_n = \left\{(x_1, \dots, x_n) \,\middle|\, \left|-\frac{1}{n} \log p(x_1, \dots, x_n) - H\right| \leq \varepsilon\right\} = \{(x_1, \dots, x_n) \mid \text{condition (ii) holds}\}$$

For the converse,

$$\mathbb{P}\left(\left|\frac{1}{n} \log p(x_1, \dots, x_n) - H\right| < \varepsilon\right) \geq \mathbb{P}(T_n) \to 1$$

□

4.3. Shannon's first coding theorem

Theorem. Let $X_1, X_2, \dots$ be a source satisfying the asymptotic equipartition property with constant H. Then this source has information rate H.

Proof. Let $\varepsilon > 0$, and let $T_n \subseteq \mathcal{A}^n$ be typical sets. Then, for all $n \geq n_0(\varepsilon)$, for all $(x_1, \dots, x_n) \in T_n$ we have $p(x_1, \dots, x_n) \geq 2^{-n(H+\varepsilon)}$. Therefore, $1 \geq \mathbb{P}(T_n) \geq 2^{-n(H+\varepsilon)} \cdot |T_n|$, giving $\frac{1}{n} \log |T_n| \leq H + \varepsilon$. Taking $A_n = T_n$ in the definition of reliable encoding shows that the source is reliably encodable at rate $H + \varepsilon$.

Conversely, if $H = 0$ the proof concludes, so we may assume $H > 0$. Let $0 < \varepsilon < \frac{H}{2}$, and suppose that the source is reliably encodable at rate $H - 2\varepsilon$ with sets $A_n \subseteq \mathcal{A}^n$. Let $T_n \subseteq \mathcal{A}^n$ be typical sets. Then, for all $(x_1, \dots, x_n) \in T_n$, $p(x_1, \dots, x_n) \leq 2^{-n(H-\varepsilon)}$, so $\mathbb{P}(A_n \cap T_n) \leq 2^{-n(H-\varepsilon)}|A_n|$, giving

$$\frac{1}{n} \log \mathbb{P}(A_n \cap T_n) \leq -(H - \varepsilon) + \frac{1}{n} \log |A_n| \to -(H - \varepsilon) + (H - 2\varepsilon) = -\varepsilon$$

Then, $\log \mathbb{P}(A_n \cap T_n) \to -\infty$, so $\mathbb{P}(A_n \cap T_n) \to 0$. But $\mathbb{P}(T_n) \leq \mathbb{P}(A_n \cap T_n) + \mathbb{P}(\mathcal{A}^n \setminus A_n) \to 0 + 0$, contradicting typicality. So we cannot reliably encode at rate $H - \varepsilon$, so the information rate is at least H. □

Corollary. A Bernoulli source $X_1, X_2, \dots$ has information rate $H(X_1)$.

Proof. In a previous example we showed that for a Bernoulli source, $-\frac{1}{n} \log p(X_1, \dots, X_n) \xrightarrow{\mathbb{P}} H(X_1)$. So the asymptotic equipartition property holds with $H = H(X_1)$, giving the result by Shannon's first coding theorem. □

Remark. The asymptotic equipartition property is useful for noiseless coding. We can encode the typical sequences using a block code, and encode the atypical sequences arbitrarily.

Many sources, which are not necessarily Bernoulli, also satisfy the property. Under suitable hypotheses, the sequence $\frac{1}{n} H(X_1, \dots, X_n)$ is decreasing, and the asymptotic equipartition property is satisfied with constant $H = \lim_{n \to \infty} \frac{1}{n} H(X_1, \dots, X_n)$.

4.4. Capacity

Consider a communication channel with input alphabet $\mathcal{A}$ and output alphabet $\mathcal{B}$. Recall the following definitions. A *code* of length n is a subset $C \subseteq \mathcal{A}^n$. The *error rate* is

$$\hat{e}(C) = \max_{c \in C} \mathbb{P}(\text{error} \mid c \text{ sent})$$

The *information rate* is $\rho(C) = \frac{\log |C|}{n}$. A channel can *transmit reliably at rate R* if there exist codes $C_1, C_2, \dots$ where C_n has length n such that $\lim_{n \to \infty} \rho(C_n) = R$ and $\lim_{n \to \infty} \hat{e}(C_n) = 0$.

The *(operational) capacity* of a channel is the supremum of all rates at which it can transmit reliably.

Suppose we are given a source with information rate r bits per second that emits symbols at a rate of s symbols per second. Suppose we also have a channel with capacity R bits per transmission that transmits symbols at a rate of S transmissions per second. Usually, information theorists take $S = s = 1$. We will show that reliable encoding and transmission is possible if and only if $rs \leq RS$.

We will now compute the capacity of the binary symmetric channel with error probability p.

Proposition. A binary symmetric channel with error probability $p < \frac{1}{4}$ has nonzero capacity.

Proof. Let δ be such that $2p < \delta < \frac{1}{2}$. We claim that we can reliably transmit at rate $R = 1 - H(\delta) > 0$. Let C_n be a code of length n, and suppose it has minimum distance $\lfloor n\delta \rfloor$ of maximal size. Then, by the GSV bound,

$$|C_n| = A(n, \lfloor n\delta \rfloor) \geq 2^{-n(1-H(\delta))} = 2^{nR}$$

Replacing C_n with a subcode if necessary, we can assume $|C_n| = \lfloor 2^{nR} \rfloor$, with minimum distance at least $\lfloor n\delta \rfloor$. Using minimum distance decoding,

$$\begin{aligned} \hat{e}(C_n) &\leq \mathbb{P}\left(\text{in } n \text{ uses, the channel makes at least } \left\lfloor \frac{\lfloor n\delta \rfloor - 1}{2} \right\rfloor \text{ errors}\right) \\ &\leq \mathbb{P}\left(\text{in } n \text{ uses, the channel makes at least } \left\lfloor \frac{n\delta - 1}{2} \right\rfloor \text{ errors}\right) \end{aligned}$$

Let $\varepsilon > 0$ be such that $p + \varepsilon < \frac{\delta}{2}$. Then, for n sufficiently large, $\frac{n\delta - 1}{2} = n\left(\frac{\delta}{2} - \frac{1}{2n}\right) > n(p + \varepsilon)$. Hence, $\hat{e}(C_n) \leq \mathbb{P}\left(\text{in } n \text{ uses, the channel makes at least } n(p+\varepsilon) \text{ errors}\right)$. We show that this value converges to zero as $n \to \infty$ using the next lemma. □

Lemma. Let $\varepsilon > 0$. A binary symmetric channel with error probability p is used to transmit n digits. Then,

$$\lim_{n \to \infty} \mathbb{P}\left(\text{in } n \text{ uses, the channel makes at least } n(p+\varepsilon) \text{ errors}\right) = 0$$

Proof. Consider random variables $U_i = \mathbb{1}[\text{the } i\text{th digit is mistransmitted}]$. The U_i are independent and identically distributed with $\mathbb{P}(U_i = 1) = p$. In particular, $\mathbb{E}[U_i] = p$. Therefore, the probability that the channel makes at least $n(p + \varepsilon)$ errors is

$$\mathbb{P}\left(\sum_{i=1}^{n} U_i \geq n(p+\varepsilon)\right) \leq \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} U_i - p\right| \geq \varepsilon\right)$$

so the result holds by the weak law of large numbers. □

4.5. Conditional entropy

Definition. Let X, Y be random variables taking values in alphabets $\mathcal{A}, \mathcal{B}$ respectively. Then, the *conditional entropy* is defined by

$$H(X \mid Y = y) = -\sum_{x \in \mathcal{A}} \mathbb{P}(X = x \mid Y = y) \log \mathbb{P}(X = x \mid Y = y)$$

and

$$H(X \mid Y) = \sum_{y \in \mathcal{B}} \mathbb{P}(Y = y) H(X \mid Y = y)$$

Note that $H(X \mid Y) \geq 0$.

Lemma. $H(X, Y) = H(X \mid Y) + H(Y)$.

Proof.

$$\begin{aligned}
H(X \mid Y) &= -\sum_{y \in \mathcal{B}} \sum_{x \in \mathcal{A}} \mathbb{P}(X = x \mid Y = y) \mathbb{P}(Y = y) \log(\mathbb{P}(X = x \mid Y = y)) \\
&= -\sum_{y \in \mathcal{B}} \sum_{x \in \mathcal{A}} \mathbb{P}(X = x \mid Y = y) \mathbb{P}(Y = y) \log\left(\frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}\right) \\
&= -\sum_{y \in \mathcal{B}} \sum_{x \in \mathcal{A}} \mathbb{P}(X = x, Y = y)(\log \mathbb{P}(X = x, Y = y) - \log \mathbb{P}(Y = y)) \\
&= -\sum_{y \in \mathcal{B}} \sum_{x \in \mathcal{A}} \mathbb{P}(X = x, Y = y) \log \mathbb{P}(X = x, Y = y) \\
&+ \sum_{y \in \mathcal{B}} \sum_{x \in \mathcal{A}} \mathbb{P}(X = x, Y = y) \log \mathbb{P}(Y = y) \\
&= -\sum_{y \in \mathcal{B}} \sum_{x \in \mathcal{A}} \mathbb{P}(X = x, Y = y) \log \mathbb{P}(X = x, Y = y) \\
&+ \sum_{y \in \mathcal{B}} \mathbb{P}(Y = y) \log \mathbb{P}(Y = y) \\
&= H(X, Y) - H(Y)
\end{aligned}$$

□

Example. Let X be a uniform random variable on $\{1, \dots, 6\}$ modelling a dice roll, and Y is defined to be zero if X is even, and one if X is odd. Then, $H(X, Y) = H(X) = \log 6$ and $H(Y) = \log 2$. Therefore, $H(X \mid Y) = \log 3$ and $H(Y \mid X) = 0$.

Corollary. $H(X \mid Y) \leq H(X)$, with equality if and only if X and Y are independent.

Proof. Combine this result with the fact that $H(X, Y) \leq H(X) + H(Y)$ where equality holds if and only if $H(X), H(Y)$ are independent. □

Now, replace random variables X and Y with random vectors $X^{(r)} = (X_1, \dots, X_r)$ and $Y^{(s)} = (Y_1, \dots, Y_s)$. Similarly, we can define $H(X_1, \dots, X_r \mid Y_1, \dots, Y_s) = H(X^{(r)} \mid Y^{(s)})$. Note that $H(X, Y \mid Z)$ is the entropy of X and Y combined, given the value of Z, and is not the entropy of X, together with Y given Z.

Lemma. Let X, Y, Z be random variables. Then, $H(X \mid Y) \leq H(X \mid Y, Z) + H(Z)$.

Proof. Expand $H(X, Y, Z)$ in two ways.

$$H(Z \mid X, Y) + \underbrace{H(X \mid Y) + H(Y)}_{H(X,Y)} = H(X, Y, Z) = H(X \mid Y, Z) + \underbrace{H(Z \mid Y) + H(Y)}_{H(Y,Z)}$$

Since $H(Z \mid X, Y) \geq 0$, we have

$$H(X \mid Y) \leq H(X \mid Y, Z) + H(Z \mid Y) \leq H(X \mid Y, Z) + H(Z)$$

□

Proposition (Fano's inequality)**.** Let X, Y be random variables taking values in $\mathcal{A}$. Let $|\mathcal{A}| = m$, and let $p = \mathbb{P}(X \neq Y)$. Then $H(X \mid Y) \leq H(p) + p\log(m-1)$.

Proof. Define Z to be zero if $X = Y$ and one if $X \neq Y$. Then, $\mathbb{P}(Z = 0) = \mathbb{P}(X = Y) = 1 - p$, and $\mathbb{P}(Z = 1) = \mathbb{P}(X \neq Y) = p$. Hence, $H(Z) = H(p)$. Applying the previous lemma, $H(X \mid Y) \leq H(X \mid Y, Z) + H(p)$, so it suffices to show $H(X \mid Y, Z) \leq p\log(m-1)$.

Since $Z = 0$ implies $X = Y$, $H(X \mid Y = y, Z = 0) = 0$. There are $m-1$ remaining possibilities for X. Hence, $H(X \mid Y = y, Z = 1) \leq \log(m-1)$.

$$\begin{aligned}
H(X \mid Y, Z) &= \sum_{y \in \mathcal{A}} \sum_{z \in \{0,1\}} \mathbb{P}(Y = y, Z = z) H(X \mid Y = y, Z = z) \\
&\leq \sum_{y \in \mathcal{A}} \mathbb{P}(Y = y, Z = 1) \log(m-1) \\
&= \mathbb{P}(Z = 1) \log(m-1) \\
&= p\log(m-1)
\end{aligned}$$

as required. □

Let X be a random variable describing the input to a channel and Y be a random variable describing the output of the channel. $H(p)$ provides the information required to decide whether an error has occurred, and $p\log(m-1)$ gives the information needed to resolve that error in the worst possible case.

4.6. Shannon's second coding theorem

Definition. Let X, Y be random variables taking values in $\mathcal{A}$. The *mutual information* is $I(X;Y) = H(X) - H(X \mid Y)$.

This is nonnegative, as $I(X;Y) = H(X) + H(Y) - H(X,Y) \geq 0$. Equality holds if and only if X, Y are independent. Clearly, $I(X;Y) = I(Y;X)$.

Definition. Consider a discrete memoryless channel with input alphabet $\mathcal{A}$ of size m and output alphabet $\mathcal{B}$. Let X be a random variable taking values in $\mathcal{A}$, used as the input to this channel. Let Y be the random variable output by the channel, depending on X and the channel matrix. The *information capacity* of the channel is $\max_X I(X;Y)$.

The maximum is taken over all discrete random variables X taking values in $\mathcal{A}$, or equivalently. This maximum is attained since I is continuous and the space

$$\left\{(p_1, \dots, p_m) \in \mathbb{R}^m \,\middle|\, p_i \geq 0, \sum_{i=1}^m p_i = 1\right\}$$

is compact. The information capacity depends only on the channel matrix.

Theorem. For a discrete memoryless channel, the (operational) capacity is equal to the information capacity.

We prove that the operational capacity is at most the information capacity in general, and we will prove the other inequality for the special case of the binary symmetric channel.

Example. Assuming this result holds, we compute the capacity of certain specific channels.

(i) Consider the binary symmetric channel with error probability p, input X, and output Y. Let $\mathbb{P}(X=0) = \alpha, \mathbb{P}(X=1) = 1-\alpha$, so $\mathbb{P}(Y=0) = (1-p)\alpha p(1-\alpha), \mathbb{P}(Y=1) = (1-p)(1-\alpha) + p\alpha$. Then, as $H(Y \mid X) = \mathbb{P}(X=0)H(p) + \mathbb{P}(X=1)H(p)$,

$$\begin{aligned} C &= \max_\alpha I(X;Y) = \max_\alpha [H(Y) - H(Y \mid X)] \\ &= \max_\alpha [H(\alpha(1-p) + (1-\alpha)p) - H(p)] = 1 - H(p) \end{aligned}$$

with the maximum attained at $\alpha = \frac{1}{2}$. Hence, the capacity of the binary symmetric channel is $C = 1 + p\log p + (1-p)\log(1-p)$. If $p = 0$ or $p = 1$, $C = 1$. If $p = \frac{1}{2}$, $C = 0$. Note that $I(X;Y) = I(Y;X)$; we can choose which to calculate for convenience.

(ii) Consider the binary erasure channel with erasure probability p, input X, and output Y. Let $\mathbb{P}(X=0) = \alpha, \mathbb{P}(X=1) = 1-\alpha$, so $\mathbb{P}(Y=0) = (1-p)\alpha, \mathbb{P}(Y=1) = (1-p)(1-\alpha), \mathbb{P}(Y=\star) = p$. We obtain

$$H(X \mid Y=0) = 0; \quad H(X \mid Y=1) = 0; \quad H(X \mid Y=\star) = H(\alpha)$$

Therefore, $H(X \mid Y) = pH(\alpha)$, giving

$$\begin{aligned} C &= \max_\alpha I(X;Y) = \max_\alpha [H(X) - H(X \mid Y)] \\ &= \max_\alpha [H(\alpha) - pH(\alpha)] = (1-p) \max_\alpha H(\alpha) = 1 - p \end{aligned}$$

with maximum attained at $\alpha = \frac{1}{2}$.

We will now model using a channel n times as the *nth extension*, replacing $\mathcal{A}$ with $\mathcal{A}^n$ and $\mathcal{B}$ with $\mathcal{B}^n$, and use the channel matrix defined by

$$\mathbb{P}(y_1 \dots y_n \text{ received} \mid x_1 \dots x_n \text{ sent}) = \prod_{i=1}^n \mathbb{P}(y_i \mid x_i)$$

Lemma. Consider a discrete memoryless channel with information capacity C. Then, its nth extension has information capacity nC.

Proof. Let $X_1, \dots, X_n$ be the input producing an output $Y_1, \dots, Y_n$. Since the channel is memoryless,

$$H(Y_1, \dots, Y_n \mid X_1, \dots, X_n) = \sum_{i=1}^n H(Y_i \mid X_1, \dots, X_n) = \sum_{i=1}^n H(Y_i \mid X_i)$$

Therefore,

$$\begin{aligned} I(X_1, \dots, X_n; Y_1, \dots, Y_n) &= H(Y_1, \dots, Y_n) - H(Y_1, \dots, Y_n \mid X_1, \dots, X_n) \\ &\leq \sum_{i=1}^n H(Y_i) - \sum_{i=1}^n H(Y_i \mid X_i) \\ &= \sum_{i=1}^n [H(Y_i) - H(Y_i \mid X_i)] \\ &= \sum_{i=1}^n I(X_i; Y_i) \leq nC \end{aligned}$$

Equality is attained by taking $X_1, \dots, X_n$ independent and identically distributed such that $I(X_i; Y_i) = C$. Indeed, if $X_1, \dots, X_n$ are independent, then so are $Y_1, \dots, Y_n$, so $H(Y_1, \dots, Y_n) = \sum_{i=1}^n H(Y_i)$. Therefore,

$$\max_{X_1, \dots, X_n} I(X_1, \dots, X_n; Y_1, \dots, Y_n) = nC$$

as required. □

We now prove part of Shannon's second coding theorem, that the operational capacity is at most the information capacity for a discrete memoryless channel.

Proof. Let C be the information capacity. Suppose reliable transmission is possible at a rate $R > C$. Then, there is a sequence of codes $(C_n)_{n\geq 1}$ where C_n has length n and size $\lfloor 2^{nR} \rfloor$, such that $\lim_{n\to\infty} \rho(C_n) = R$ and $\lim_{n\to\infty} \hat{e}(C_n) = 0$.

Recall that $\hat{e}(C_n) = \max_{c\in C_n} \mathbb{P}\left(\text{error} \mid c \text{ sent}\right)$. Define the *average error rate* $e(C)$ by $e(C) = \frac{1}{|C_n|} \sum_{c\in C} \mathbb{P}\left(\text{error} \mid c \text{ sent}\right)$. Note that $e(C_n) \leq \hat{e}(C_n)$. As $\hat{e}(C_n) \to 0$, we also have $e(C_n) \to 0$.

Consider an input random variable X distributed uniformly over C_n. Let Y be the output given by X and the channel matrix. Then $e(C_n) = \mathbb{P}\left(X \neq Y\right) = p$. Hence, $H(X) = \log|C_n| = \log\lfloor 2^{nR} \rfloor \geq nR - 1$ for sufficiently large n. Also, by Fano's inequality, $H(X \mid Y) \leq H(p) + p\log(|C_n| - 1) \leq 1 + pnR$.

Recall that $I(X;Y) = H(X) - H(X \mid Y)$. By the previous lemma, $nC \geq I(X;Y)$, so

$$nC \geq nR - 1 - 1 - pnR \implies pnR \geq n(R - c) - 2 \implies p \geq \frac{n(R-C) - 2}{nR}$$

As $n \to \infty$, the right hand side converges to $\frac{R-C}{R} > 0$. This contradicts the fact that $p = e(C_n) \to 0$. Hence, we cannot transmit reliably at any rate which exceeds C, hence the capacity is at most C. □

To complete the proof of Shannon's second coding theorem for the binary symmetric channel with error probability p, we prove that the operational capacity is at least $1 - H(p)$.

Proposition. Consider a binary symmetric channel with error probability p, and let $R < 1 - H(p)$. Then there exists a sequence of codes $(C_n)_{n\geq 1}$ with C_n of length n and size $\lfloor 2^{nR} \rfloor$ such that $\lim_{n\to\infty} \rho(C_n) = R$ and $\lim_{n\to\infty} e(C_n) = 0$.

Remark. This proposition deals with the average error rate, instead of the error rate $\hat{e}$.

Proof. We use the method of random coding. Without loss of generality let $p < \frac{1}{2}$. Let $\varepsilon > 0$ such that $p + \varepsilon < \frac{1}{2}$ and $R < 1 - H(p + \varepsilon)$. We use minimum distance decoding, and in the case of a tie, we make an arbitrary choice. Let $m = \lfloor 2^{nR} \rfloor$, and let $C = \{c_1, \dots, c_m\}$ be a code chosen uniformly at random from $\mathcal{C} = \{[n,m]\text{-codes}\}$, a set of size $\binom{2^n}{m}$.

Choose $1 \leq i \leq m$ uniformly at random, and send c_i through the channel, and obtain an output Y. Then, $\mathbb{P}\left(Y \text{ not decoded as } c_i\right)$ is the average value of $e(C)$ for C ranging over $\mathcal{C}$, giving $\frac{1}{|\mathcal{C}|} \sum_{C\in\mathcal{C}} e(C)$. We can choose a code $C_n \in \mathcal{C}$ such that $e(C_n) \leq \frac{1}{|\mathcal{C}|} \sum_{C\in\mathcal{C}} e(C)$. So it suffices to show $\mathbb{P}\left(Y \text{ not decoded as } c_i\right) \to 0$.

Let $r = \lfloor n(p + \varepsilon) \rfloor$. Then if $B(Y,r) \cap C = \{c_i\}$, Y is correctly decoded as c_i. Therefore,

$$\mathbb{P}\left(Y \text{ not decoded as } c_i\right) \leq \mathbb{P}\left(c_i \notin B(Y,r)\right) + \mathbb{P}\left(B(Y,r) \cap C \supsetneq \{c_i\}\right)$$

We consider the two cases separately.

In the first case with $d(c_i, Y) > r$, $\mathbb{P}\left(d(c_i, Y) > r\right)$ is the probability that the channel makes more than r errors, and hence more than $n(p+\varepsilon)$ errors. We have already shown that this converges to zero as $n \to \infty$.

In the second case with $d(c_i, Y) \leq r$, if $j \neq i$,

$$\mathbb{P}\left(c_j \in B(Y,r) \mid c_i \in B(Y,r)\right) = \frac{V(n,r)-1}{2^n - 1} \leq \frac{V(n,r)}{2^n}$$

Therefore,

$$\begin{aligned}
\mathbb{P}\left(B(Y,r) \cap C \supsetneq \{c_i\}\right) &\leq \sum_{j \neq i} \mathbb{P}\left(c_j \in B(Y,r), c_i \in B(Y,r)\right) \\
&\leq \sum_{j \neq i} \mathbb{P}\left(c_j \in B(Y,r) \mid c_i \in B(Y,r)\right) \\
&\leq (m-1)\frac{V(n,r)}{2^n} \\
&\leq \frac{mV(n,r)}{2^n} \\
&\leq 2^{nR} 2^{nH(p+\varepsilon)} 2^{-n} \\
&= 2^{n(R-(1-H(p+\varepsilon)))} \to 0
\end{aligned}$$

as required. □

Proposition. We can replace e with $\hat{e}$ in the previous result.

Proof. Let R' be such that $R < R' < 1 - H(p)$. Then, apply the previous result to R' to construct a sequence of codes $(C_n')_{n \geq 1}$ of length n and size $\lfloor 2^{nR'} \rfloor$, where $e(C_n') \to 0$. Order the codewords of C_n' by the probability of error given that the codeword was sent, and delete the worst half. This gives a code C_n with $\hat{e}(C_n) \leq 2e(C_n')$. Hence $\hat{e}(C_n) \to 0$ as $n \to \infty$. Since C_n has length n, and size $\frac{1}{2}\lfloor 2^{nR'} \rfloor = \lfloor 2^{nR'-1} \rfloor$. But $2^{nR'-1} = 2^{n(R' - \frac{1}{n})} \geq 2^{nR}$ for sufficiently large n. So we can replace C_n' with a code of smaller size $\lfloor 2^{nR} \rfloor$ and still have $\hat{e}(C_n) \to 0$ and $\rho(C_n) \to R$ as $n \to \infty$. □

Therefore, a binary symmetric channel with error probability p has operational capacity $1 - H(p)$, as we can transmit reliably at any rate $R < 1 - H(p)$, and the capacity is at most $1 - H(p)$. The result shows that codes with certain properties exist, but does not give a way to construct them.

4.7. The Kelly criterion

Let $0 < p < 1$, $u > 0$, $0 \leq w < 1$. Suppose that a coin is tossed n times in succession with probability p of obtaining a head. If a stake of k is paid ahead of a particular throw, the return is ku if the result is a head, and the return is zero if the result is a tail.

Suppose the initial bankroll is $X_0 = 1$. After n throws, the bankroll is X_n. We bet wX_n on the $(n+1)$th coin toss, retaining $(1-w)X_n$. The bankroll after the toss is

$$X_{n+1} = \begin{cases} X_n(wu + (1-w)) & (n+1)\text{th toss is a head} \\ X_n(1-w) & (n+1)\text{th toss is a tail} \end{cases}$$

Define $Y_{n+1} = \frac{X_{n+1}}{X_n}$, then the Y_i are independent and identically distributed. Then $\log Y_i$ is a sequence of independent and identically distributed random variables. Note that $\log X_n = \sum_{i=1}^n \log Y_i$.

Lemma. Let $\mu = \mathbb{E}\left[\log Y_1\right], \sigma^2 = \operatorname{Var}(\log Y_1)$. Then, if $a > 0$,

(i) $\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \log Y_i - \mu\right| \geq a\right) \leq \frac{\sigma^2}{na^2}$ by Chebyshev's inequality;

(ii) $\mathbb{P}\left(\left|\frac{\log X_n}{n} - \mu\right| \geq a\right) \leq \frac{\sigma^2}{na^2}$;

(iii) given $\varepsilon > 0$ and $\delta > 0$, there exists N such that $\mathbb{P}\left(\left|\frac{\log X_n}{n} - \mu\right| \geq \delta\right) \leq \varepsilon$ for all $n \geq N$.

Consider a single coin toss, with probability $p < 1$ of a head. Suppose that a bet of k on a head gives a payout of ku for some payout ratio $u > 0$. Suppose further that we have an initial bankroll of 1, and we bet w on heads, retaining $1-w$, for some $0 \leq w < 1$. Then, if Y is the expected fortune after the throw, $\mathbb{E}\left[\log Y\right] = p\log(1 + (u-1)w) + (1-p)\log(1-w)$. One can show that the value of $\mathbb{E}\left[\log Y\right]$ is maximised by taking $w = 0$ if $up \leq 1$, and setting $w = \frac{up-1}{u-1}$ if $up > 1$.

Let $q = 1 - p$. If $up > 1$, at the optimum value of w, we find

$$\mathbb{E}\left[\log Y\right] = p\log p + q\log q + \log u - q\log(u-1) = -H(p) + \log u - q\log(u-1)$$

Kelly's criterion is that in order to maximise profit, $\mathbb{E}\left[\log Y\right]$ should be optimised, given that we can bet arbitrarily many times.

One can show that if w is set below the optimum, the bankroll will still increase, but does so more slowly. If w is set sufficiently high, the bankroll will tend to decrease.

5. Algebraic coding theory

5.1. Linear codes

Definition. A binary code $C \subseteq \mathbb{F}_2^n$ is *linear* if $0 \in C$, and whenever $x, y \in C$, we have $x + y \in C$.

Equivalently, C is a vector subspace of $\mathbb{F}_2^n$.

Definition. The *rank* of a linear code C, denoted $\operatorname{rank} C$, is its dimension as an $\mathbb{F}_2$-vector space. A linear code of length n and rank k is called an (n, k)-code. If it has minimum distance d, it is called an (n, k, d)-code.

Let $v_1, \dots, v_k$ be a basis for C. Then $C = \left\{\sum_{i=1}^{k} \lambda_i v_i \mid \lambda_i \in \mathbb{F}_2\right\}$. The size of the code is therefore 2^k, so an (n, k)-code is an $[n, 2^k]$-code, and an (n, k, d)-code is an $[n, 2^k, d]$-code. The information rate is $\frac{k}{n}$.

Definition. The *weight* of $x \in \mathbb{F}_2^n$ is $w(x) = d(x, 0)$.

Lemma. The minimum distance of a linear code is the minimum weight of a nonzero codeword.

Proof. Let $x, y \in C$. Then, $d(x, y) = d(x + y, 0) = w(x + y)$. Observe that $x \neq y$ if and only if $x + y \neq 0$, so $d(C)$ is the minimum $w(x + y)$ for $x + y \neq 0$. □

Definition. Let $x, y \in \mathbb{F}_2^n$. Define $x \cdot y = \sum_{i=1}^{n} x_i y_i \in \mathbb{F}_2$. This is symmetric and bilinear.

There are nonzero x such that $x \cdot x = 0$.

Definition. Let $P \subseteq \mathbb{F}_2^n$. The *parity check code* defined by P is

$$C = \{x \in \mathbb{F}_2^n \mid \forall p \in P,\ p \cdot x = 0\}$$

Example. (i) $P = \{11 \dots 1\}$ gives the simple parity check code.

(ii) $P = \{1010101, 0110011, 0001111\}$ gives Hamming's original $[7, 16, 3]$-code.

(iii) C^+ and C^- are linear if C is linear.

Lemma. Every parity check code is linear.

Proof. $0 \in C$ as $p \cdot 0 = 0$. If $p \cdot x = 0$ and $p \cdot y = 0$ then $p \cdot (x + y) = 0$, so $x, y \in C$ implies $x + y \in C$. □

Definition. Let $C \subseteq \mathbb{F}_2^n$ be a linear code. The *dual code* $C^\perp$ is defined by

$$C^\perp = \{x \in \mathbb{F}_2^n \mid \forall y \in C,\ x \cdot y = 0\}$$

By definition, $C^\perp$ is a parity check code, and hence is linear. Note that $C \cap C^\perp$ may contain elements other than 0.

Lemma. $\operatorname{rank} C + \operatorname{rank} C^\perp = n$.

Proof. One can prove this by defining $C^\perp$ as an annihilator from linear algebra. A proof using coding theory is shown later. □

Corollary. Let C be a linear code. Then $(C^\perp)^\perp = C$. In particular, all linear codes are parity check codes, defined by $C^\perp$.

Proof. If $x \in C$, then $x \cdot y = 0$ for all $y \in C^\perp$ by definition, so $x \in (C^\perp)^\perp$. Then $\operatorname{rank} C = n - \operatorname{rank} C^\perp = n - (n - \operatorname{rank}(C^\perp)^\perp) = \operatorname{rank}(C^\perp)^\perp$, so $C = (C^\perp)^\perp$. □

Definition. Let C be an (n,k)-code. A *generator matrix* G for C is a $k \times n$ matrix where the rows form a basis for C. A *parity check matrix* H for C is a generator matrix for the dual code $C^\perp$, so it is an $(n-k) \times n$ matrix.

The codewords of a linear code can be viewed either as linear combinations of rows of G, or linear dependence relations between the columns of H, so $C = \{x \in \mathbb{F}_2^n \mid Hx = 0\}$.

Definition. Let C be an (n,k)-code. The *syndrome* of $x \in \mathbb{F}_2^n$ is Hx.

If we receive a word $x = c + z$ where $c \in C$ and z is the error pattern, $Hx = Hz$ as $Hc = 0$. If C is e-error correcting, we precompute Hz for all z for which $w(z) \leq e$. On receiving x, we can compute the syndrome Hx and find this entry in the table of values of Hz. If successful, we decode $c = x - z$, with $d(x,c) = w(z) \leq e$.

Definition. Codes $C_1, C_2 \subseteq \mathbb{F}_2^n$ are *equivalent* if there exists a permutation of bits that maps codewords in C_1 to codewords in C_2.

Codes are typically only considered up to equivalence.

Lemma. Every (n,k)-linear code is equivalent to one with generator matrix with block form $\begin{pmatrix} I_k & B \end{pmatrix}$ for some $k \times (n-k)$ matrix B.

Proof. Let G be a $k \times n$ generator matrix for C. Using Gaussian elimination, we can transform G into row echelon form

$$G_{ij} = \begin{cases} 0 & j < \ell(i) \\ 1 & j = \ell(i) \end{cases}$$

for some $\ell(1) < \ell(2) < \dots < \ell(k)$. Permuting the columns replaces C with an equivalent code, so without loss of generality we may assume $\ell(i) = i$. Hence,

$$G = \begin{pmatrix} 1 & & \star & \\ & \ddots & & B \\ & & 1 & \end{pmatrix}$$

Further row operations eliminate $\star$ to give G in the required form. □

A message $y \in \mathbb{F}_2^k$ viewed as a row vector can be encoded as yG. If $G = \begin{pmatrix} I_k & B \end{pmatrix}$, then $yG = (y, yB)$ where y is the message and yB is a string of check digits. We now prove the following lemma that was stated earlier.

Lemma. $\operatorname{rank} C + \operatorname{rank} C^\perp = n$.

Proof. Let C have generator matrix $G = \begin{pmatrix} I_k & B \end{pmatrix}$. G has k linearly independent columns, so there is a linear map $\gamma\colon \mathbb{F}_2^n \to \mathbb{F}_2^k$ defined by $x \mapsto Gx$ which is surjective. Its kernel is $C^\perp$. By the rank-nullity theorem, $\dim \mathbb{F}_2^n = \dim \ker \gamma + \dim \operatorname{Im} \gamma$, so $n = \operatorname{rank} C + \operatorname{rank} C^\perp$ as required. □

Lemma. An (n,k)-code with generator matrix $G = \begin{pmatrix} I_k & B \end{pmatrix}$ has parity check matrix H of the form $\begin{pmatrix} B^\intercal & I_{n-k} \end{pmatrix}$.

Proof.

$$GH^\intercal = \begin{pmatrix} I_k & B \end{pmatrix} \begin{pmatrix} B \\ I_{n-k} \end{pmatrix} = B + B = 2B = 0$$

So the rows of H generate a subcode of $C^\perp$. But $\operatorname{rank} H = n - k$, and $\operatorname{rank} C^\perp = n - k$. So $H = C^\perp$, and $C^\perp$ has generator matrix H. □

Lemma. Let C be a linear code with parity check matrix H. Then, $d(C) = d$ if and only if

(i) any $d - 1$ columns of H are linearly independent; and

(ii) a set of d columns of H are linearly dependent.

The proof is left as an exercise.

5.2. Hamming codes

Definition. Let $d \geq 1$, and let $n = 2^d - 1$. Let H be the $d \times n$ matrix with columns given by the nonzero elements of $\mathbb{F}_2^d$. The *Hamming $(n, n-d)$-linear code* is the code with parity check matrix H.

Lemma. The Hamming $(n, n-d)$-code C has minimum distance $d(C) = 3$, and is a perfect 1-error correcting code.

Proof. Any two columns of H are linearly independent, but there are three linearly dependent columns. Hence, $d(C) = 3$. Hence, C is $\left\lfloor \frac{3-1}{2} \right\rfloor = 1$-error correcting. A perfect code is one such that $|C| = \frac{2^n}{V(n,e)}$. In this case, $n = 2^d - 1$ and $e = 1$, so $\frac{2^n}{1+2^d-1} = 2^{n-d} = |C|$ as required. □

5.3. Reed–Muller codes

Let $X = \{p_1, \dots, p_n\}$ be a set of size n. There is a correspondence between the power set $\mathcal{P}(X)$ and $\mathbb{F}_2^n$.

$$\mathcal{P}(X) \xrightarrow{A \mapsto \mathbb{1}_A} \{f : X \to \mathbb{F}_2\} \xrightarrow{f \mapsto (f(p_1), \dots, f(p_n))} \mathbb{F}_2^n$$

The *symmetric difference* of two sets A, B is defined to be $A \bigtriangleup B = A \setminus B \cup B \setminus A$, which corresponds to vector addition in $\mathbb{F}_2^n$. Intersection $A \cap B$ corresponds to the *wedge product* $x \wedge y = (x_1 y_1, \dots, x_n y_n)$.

Let $X = \mathbb{F}_2^d$, so $n = 2^d - |X|$. Let $v_0 = (1, \dots, 1)$, and let $v_i = \mathbb{1}_{H_i}$ where $H_i = \{p \in X \mid p_i = 0\}$ is a coordinate hyperplane.

Definition. Let $0 \leq r \leq d$. The *Reed–Muller code* $RM(d, r)$ of *order* r and length 2^d is the linear code spanned by v_0 and all wedge products of at most r of the the v_i for $1 \leq i \leq d$.

By convention, the empty wedge product is v_0.

Example. Let $d = 3$, and let $X = \mathbb{F}_2^3 = \{p_1, \dots, p_8\}$ in binary order.

X	000	001	010	011	100	101	110	111
v_0	1	1	1	1	1	1	1	1
v_1	1	1	1	1	0	0	0	0
v_2	1	1	0	0	1	1	0	0
v_3	1	0	1	0	1	0	1	0
$v_1 \wedge v_2$	1	1	0	0	0	0	0	0
$v_2 \wedge v_3$	1	0	0	0	1	0	0	0
$v_1 \wedge v_3$	1	0	1	0	0	0	0	0
$v_1 \wedge v_2 \wedge v_3$	1	0	0	0	0	0	0	0

A generator matrix for Hamming's original code is a submatrix in the top-right corner.

$RM(3, 0)$ is spanned by v_0, and is hence the repetition code of length 8. $RM(3, 1)$ is spanned by v_0, v_1, v_2, v_3, which is equivalent to a parity check extension of Hamming's original $(7, 4)$-code. $RM(3, 2)$ is an $(8, 7)$-code, and can be shown to be equivalent to a simple parity check code of length 8. $RM(3, 3)$ is the trivial code $\mathbb{F}_2^8$ of length 8.

Theorem. (i) The vectors $v_{i_1} \wedge \dots \wedge v_{i_s}$ for $i_1 < \dots < i_s$ and $0 \leq s \leq d$ form a basis for $\mathbb{F}_2^n$.

(ii) The rank of $RM(d, r)$ is $\sum_{s=0}^{r} \binom{d}{s}$.

Proof. Part (i). There are $\sum_{s=0}^{d} \binom{d}{s} = 2^d = n$ vectors listed, so it suffices to show they are a spanning set, or equivalently $RM(d, d)$ is the trivial code. Let $p \in X$, and let y_i be v_i if $p_i = 0$ and $v_0 + v_i$ if $p_i = 1$. Then $\mathbb{1}_{\{p\}} = y_1 \wedge \dots \wedge y_d$. Expanding this using the distributive law, $\mathbb{1}_{\{p\}} \in RM(d, d)$. But the set of $\mathbb{1}_{\{p\}}$ for $p \in X$ spans $\mathbb{F}_2^n$, as required.

Part (ii). $RM(d,r)$ is spanned by $v_{i_1} \wedge \dots \wedge v_{i_s}$ where $i_1 < \dots < i_s$ and $0 \leq s \leq r$. Since these are linearly independent, the rank of $RM(d,r)$ is the number of such vectors, which is $\sum_{s=0}^{d} \binom{d}{s}$. □

Definition. Let C_1, C_2 be linear codes of length n where $C_2 \subseteq C_1$. The *bar product* is $C_1 \mid C_2 = \{(x \mid x+y) \mid x \in C_1, y \in C_2\}$.

This is a linear code of length $2n$.

Lemma. (i) $\operatorname{rank}(C_1 \mid C_2) = \operatorname{rank} C_1 + \operatorname{rank} C_2$.

(ii) $d(C_1 \mid C_2) = \min\{2d(C_1), d(C_2)\}$.

Proof. Part (i). If C_1 has basis $x_1, \dots, x_k$ and C_2 has basis $y_1, \dots, y_\ell$, then $C_1 \mid C_2$ has basis

$$\{(x_i \mid x_i) \mid 1 \leq i \leq k\} \cup \{(0 \mid y_i) \mid 1 \leq i \leq \ell\}$$

Part (ii). Let $0 \neq (x \mid x+y) \in C_1 \mid C_2$. If $y \neq 0$, then $w(x \mid x+y) = w(x) + w(x+y) \geq w(y) \geq d(C_2)$. If $y = 0$, then $w(x \mid x+y) = w(x \mid x) = 2w(x) \geq 2d(C_1)$. Hence, $d(C_1 \mid C_2) \geq \min\{2d(C_1), d(C_2)\}$.

There is a nonzero $x \in C_1$ with $w(x) = d(C_1)$, so $d(C_1 \mid C_2) \leq w(x \mid x) = 2d(C_1)$. There is a nonzero $y \in C_2$ with $w(y) = d(C_2)$, giving $d(C_1 \mid C_2) \leq w(0 \mid 0+y) = d(C_2)$, giving the other inequality as required. □

Theorem. (i) $RM(d,r) = RM(d-1,r) \mid RM(d-1,r-1)$ for $0 < r < d$.

(ii) $RM(d,r)$ has minimum distance 2^{d-r} for all r.

Proof. Part (i). Exercise.

Part (ii). If $r = 0$, then $RM(d,r)$ is the repetition code of length 2^d, which has minimum distance 2^d. If $r = d$, $RM(d,r)$ is the trivial code of length 2^d, which has minimum distance $1 = 2^{d-d}$. We prove the remaining cases by induction on d. From part (i), $RM(d,r) = RM(d-1,r) \mid RM(d-1,r-1)$. By induction, the minimum distance of $RM(d-1,r)$ is 2^{d-1-r} and the minimum distance of $RM(d-1,r-1)$ is 2^{d-r}. By part (ii) of the previous lemma, the minimum distance of $RM(d,r)$ is $\min\{2 \cdot 2^{d-1-r}, 2^{d-r}\} = 2^{d-r}$. □

5.4. Cyclic codes

If F is a field and $f \in F[X]$, ${}^{F[X]}\!/_{(f)}$ is in bijection with F^n where $n = \deg f$, since ${}^{F[X]}\!/_{(f)}$ is represented by the set of functions of degree less than $\deg f$.

Definition. A linear code $C \subseteq \mathbb{F}_2^n$ is *cyclic* if

$$(a_0, a_1, \dots, a_{n-1}) \in C \implies (a_{n-1}, a_0, \dots, a_{n-2}) \in C$$

We identify ${}^{\mathbb{F}_2[X]}\big/_{(X^n\,-\,1)}$ with $\mathbb{F}_2^n$, letting $\pi(a_0+a_1X+\dots+a_{n-1}X^{n-1}) = (a_0, a_1, \dots, a_{n-1})$.

Lemma. A code $C \subseteq \mathbb{F}_2^n$ is cyclic if and only if $\pi(\mathcal{C}) = C$ satisfies

(i) $0 \in \mathcal{C}$;

(ii) $f, g \in \mathcal{C}$ implies $f + g \in \mathcal{C}$;

(iii) $f \in \mathbb{F}_2[X], g \in \mathcal{C}$ implies $fg \in \mathcal{C}$.

Equivalently, $\mathcal{C}$ is an ideal of ${}^{\mathbb{F}_2[X]}\big/_{(X^n\,-\,1)}$.

Proof. If $g(X) = a_0 + a_1X + \dots + a_{n-1}X^{n-1}$, multiplication by X gives $Xg(X) = a_{n-1} + a_0X + \dots + a_{n-2}X^{n-1}$. So $\mathcal{C}$ is cyclic if and only if (i) and (ii) hold and $g(X) \in C$ implies $Xg(X) \in C$. Linearity then gives (iii). □

We will identify C with $\mathcal{C}$. The cyclic codes of length n correspond to ideals in ${}^{\mathbb{F}_2[X]}\big/_{(X^n\,-\,1)}$. Such ideals correspond to ideals of $\mathbb{F}_2[X]$ that contain $X^n - 1$. Since $\mathbb{F}_2[X]$ is a principal ideal domain, these ideals correspond to polynomials $g(X) \in \mathbb{F}_2[X]$ dividing $X^n - 1$.

Theorem. Let $C \trianglelefteq {}^{\mathbb{F}_2[X]}\big/_{(X^n\,-\,1)}$ be a cyclic code. Then, there exists a unique *generator* polynomial $g(X) \in \mathbb{F}_2[X]$ such that

(i) $C = (g)$;

(ii) $g(X) \mid X^n - 1$.

In particular, $p(X) \in \mathbb{F}_2[X]$ represents a codeword if and only if $g \mid p$.

Proof. Let $g(X) \in \mathbb{F}_2[X]$ be the polynomial of smallest degree that represents a nonzero codeword of C. Note that $\deg g < n$. Since C is cyclic, $(g) \subseteq C$. Now let $p(X) \in \mathbb{F}_2[X]$ represent a codeword. By the division algorithm, $p = qg + r$ for $q, r \in \mathbb{F}_2[X]$ where $\deg r < \deg g$. Then, $r = p - qg \in C$ as C is an ideal. But $\deg r < \deg g$, so $r = 0$. Hence, $g \mid p$. For part (ii), let $p(X) = X^n - 1$, giving $g \mid X^n - 1$.

Now we show uniqueness. Suppose $C = (g_1) = (g_2)$. Then $g_1 \mid g_2$ and $g_2 \mid g_1$. So $g_1 = cg_2$ where $c \in \mathbb{F}_2^\star$, so $c = 1$. □

Lemma. Let C be a cyclic code of length n with generator $g(X) = a_0 + a_1X + \dots + a_kX^k$ with $a_k \neq 0$. Then C has basis $\{g, Xg, X^2g, \dots, X^{n-k-1}g\}$. In particular, $\operatorname{rank} C = n - k$.

Proof. Exercise. □

Corollary. Let C be a cyclic code of length n with generator $g(X) = a_0 + a_1X + \dots + a_kX^k$ with $a_k \neq 0$. Then, a generator matrix for C is given by

$$G = \begin{pmatrix} a_0 & a_1 & a_2 & \cdots & a_k & 0 & 0 & \cdots & 0 \\ 0 & a_0 & a_1 & \cdots & a_{k-1} & a_k & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & a_0 & a_1 & \cdots & a_k \end{pmatrix}$$

This is an $(n-k) \times n$ matrix.

Definition. Let g be a generator for C. The *parity check polynomial* is the polynomial h such that $g(X)h(X) = X^n - 1$.

Corollary. Writing $h(X) = b_0 + b_1X + \dots + b_{n-k}X^{n-k}$, the parity check matrix is

$$H = \begin{pmatrix} b_{n-k} & b_{n-k-1} & b_{n-k-2} & \cdots & b_1 & b_0 & 0 & 0 & \cdots & 0 \\ 0 & b_{n-k} & b_{n-k-1} & \cdots & b_2 & b_1 & b_0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & b_{n-k} & b_{n-k-1} & b_{n-k-2} & \cdots & b_0 \end{pmatrix}$$

which is a $k \times n$ matrix.

Proof. One can check that the inner product of the ith row of the generator matrix and the jth row of the parity check matrix is the coefficient of $X^{n-k-i+j}$ in $g(X)h(X) = X^n - 1$. Since $1 \leq i \leq n-k$ and $1 \leq j \leq k$, $0 < n-k-i+j < n$, and such coefficients are zero. Hence, the rows of G are orthogonal to the rows of H. Note that as $b_{n-k} \neq 0$, $\operatorname{rank} H = k = \operatorname{rank} C^\perp$, so H is the parity check matrix. □

Remark. Given a polynomial $f(X) = \sum_{i=0}^m f_iX_i$ of degree m, the *reverse* polynomial is $\check{f}(X) = f_n + f_{n-1}X + \dots + f_0X^M = X^m f\left(\frac{1}{X}\right)$. The cyclic code generated by $\check{h}$ is the dual code $C^\perp$.

Lemma. If n is odd, $X^n - 1 = f_1(X) \dots f_t(X)$ where the $f_i(X)$ are distinct irreducible polynomials in $\mathbb{F}_2[X]$. Thus, there are 2^t cyclic codes of length n.

This is false if n is even, for instance, $X^2 - 1 = (X-1)^2$. The proof follows from Galois theory.

5.5. BCH codes

Recall that if p is a prime, $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ is a field, and if $f(X) \in \mathbb{F}_p[X]$ is irreducible, the quotient $K = \mathbb{F}_p[X]/(f)$ is a field and has order $p^{\deg f}$. Moreover, any finite field arises in this way.

If $q = p^\alpha$ is a prime power where $\alpha \geq 1$, there exists a unique field $\mathbb{F}_q$ of order q, up to isomorphism. Note that $\mathbb{F}_q \not\simeq \mathbb{Z}/q\mathbb{Z}$ if $\alpha > 1$. The multiplicative group $\mathbb{F}_q^\times$ is cyclic; there exists $\beta \in \mathbb{F}_q$ such that $\mathbb{F}_q^\times = \langle\beta\rangle = \{1, \beta, \dots, \beta^{q-2}\}$. Such a β is called a *primitive element*.

Let n be an odd integer, and let $r \geq 1$ such that $2^r \equiv 1 \bmod n$, which always exists as 2 is coprime to n. Let $K = \mathbb{F}_{2^r}$, and define $\boldsymbol{\mu}_n(K) = \{x \in K \mid x^n = 1\} \leq K^\times$, which is a cyclic group. Since $n \mid (2^r - 1) = |K^\times|$, $\boldsymbol{\mu}_n(K)$ is the cyclic group of order n. Hence, $\boldsymbol{\mu}_n(K) = \left\{1, \alpha, \alpha^2, \dots, \alpha^{n-1}\right\}$ for some primitive nth root of unity $\alpha \in K$.

Definition. The cyclic code of length n with *defining set* $A \subseteq \boldsymbol{\mu}_n(K)$ is the code

$$C = \left\{ f(X) \in {}^{\mathbb{F}_2[X]}\!\big/_{(X^n - 1)} \,\middle|\, \forall a \in A,\ f(a) = 0 \right\}$$

The generator polynomial $g(X)$ is the nonzero polynomial of least degree such that $g(a) = 0$ for all $a \in A$. Equivalently, g is the least common multiple of the minimal polynomials of the elements of A.

Definition. The cyclic code of length n with defining set $\left\{\alpha, \alpha^2, \dots, \alpha^{\delta-1}\right\}$ is a *BCH code* with *design distance* δ.

Theorem. A BCH code C with design distance δ has minimum distance $d(C) \geq \delta$.

This proof needs the following result.

Lemma. The Vandermonde matrix satisfies

$$\det \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_n \\ x_1^2 & x_2^2 & x_3^2 & \cdots & x_n^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^{n-1} & x_2^{n-1} & x_3^{n-1} & \cdots & x_n^{n-1} \end{pmatrix} = \prod_{1 \leq j < i \leq n} (x_i - x_j)$$

Proof of theorem. Consider

$$H = \begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{n-1} \\ 1 & \alpha^2 & \alpha^4 & \cdots & \alpha^{2(n-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \alpha^{\delta-1} & \alpha^{2(\delta-1)} & \cdots & \alpha^{(\delta-1)(n-1)} \end{pmatrix}$$

This is a $(\delta - 1) \times n$ matrix. Any collection of $(\delta - 1)$ columns is independent as it forms a Vandermonde matrix. But any codeword of C is a dependence relation between the columns of H. Hence every nonzero codeword has weight at least δ, giving $d(C) \geq \delta$. □

Note that H in the proof above is not a parity check matrix, as its entries do not lie in $\mathbb{F}_2$.

Let C be a cyclic code with defining set $\left\{\alpha, \alpha^2, \dots, \alpha^{\delta-1}\right\}$ where $\alpha \in K$ is a primitive nth root of unity. Its minimum distance is at least δ, so we should be able to correct $t = \left\lfloor \frac{\delta-1}{2} \right\rfloor$ errors. Suppose we send $c \in C$ through the channel, and receive $r = c + e$ where e is the error pattern with at most t nonzero errors. Note that r, c, e correspond to polynomials $r(X), c(X), e(X)$, and $c(\alpha^j) = 0$ for $j \in \{1, \dots, \delta - 1\}$ as c is a codeword. Hence, $r(\alpha^j) = e(\alpha^j)$.

Definition. The *error locator polynomial* of an error pattern $e \in \mathbb{F}_2^n$ is

$$\sigma(X) = \prod_{i \in \mathcal{E}} (1 - \alpha^i X) \in K[X]$$

where $\mathcal{E} = \{i \mid e_i = 1\}$.

Assuming that $\deg \sigma = |\mathcal{E}|$, where $2t + 1 \leq \delta$, we must recover σ from $r(X)$.

Theorem. Suppose $\deg \sigma = |\mathcal{E}| \leq t$ where $2t + 1 \leq \delta$. Then $\sigma(X)$ is the unique polynomial in $K[X]$ of least degree such that

(i) $\sigma(0) = 1$;

(ii) $\sigma(X) \sum_{j=1}^{2t} r(\alpha^j) X^j = \omega(X) \bmod X^{2t+1}$ for some $\omega \in K[X]$ of degree at most t.

Proof. Define $\omega(X) = -X\sigma'(X)$, called the *error co-locator*. Hence,

$$\omega(X) = \sum_{i \in \mathcal{E}} \alpha^i X \prod_{j \neq i} (1 - \alpha^j X)$$

This polynomial has $\deg \omega = \deg \sigma$. Consider the ring $K[\![X]\!]$ of formal power series. In this ring,

$$\frac{\omega(X)}{\sigma(X)} = \sum_{i \in \mathcal{E}} \frac{\alpha^i X}{1 - \alpha^i X} = \sum_{i \in \mathcal{E}} \sum_{j=1}^{\infty} (\alpha^i X)^j = \sum_{j=1}^{\infty} X^j \sum_{i \in \mathcal{E}} (\alpha^j)^i = \sum_{j=1}^{\infty} e(\alpha^j) X^j$$

Hence $\sigma(X) \sum_{j=1}^{\infty} e(\alpha^j) X^j = \omega(X)$. By definition of C, we have $c(\alpha^j) = 0$ for all $1 \leq j \leq \delta - 1$. Hence $c(\alpha^j) = 0$ for $1 \leq j \leq 2t$. As $r = c + e$, $r(\alpha^j) = e(\alpha^j)$ for all $1 \leq j \leq 2t$, hence $\sigma(X) \sum_{j=1}^{2t} r(\alpha^j) X^j = \omega(X) \bmod X^{2t+1}$. This verifies (i) and (ii) for this choice of ω, so $\deg \omega = \deg \sigma = |\mathcal{E}| \leq t$.

For uniqueness, suppose there exist $\tilde{\sigma}, \tilde{\omega}$ with the properties (i), (ii). Without loss of generality, we can assume $\deg \tilde{\sigma} \leq \deg \sigma$. $\sigma(X)$ has distinct nonzero roots, so $\omega(X) = -X\sigma'(X)$ is nonzero at these roots. Hence σ, ω are coprime polynomials. By property (ii), $\tilde{\sigma}(X)\omega(X) = \sigma(X)\tilde{\omega}(X) \bmod X^{2t+1}$. But the degrees of $\sigma, \tilde{\sigma}, \omega, \tilde{\omega}$ are at most t, so this congruence is an equality. But $\sigma(X)$ and $\omega(X)$ are coprime, so $\sigma \mid \tilde{\sigma}$, but $\deg \tilde{\sigma} \leq \deg \sigma$ by assumption, so $\tilde{\sigma} = \lambda \sigma$ for some $\lambda \in K$. By property (i), $\sigma(0) = \tilde{\sigma}(0)$ hence $\lambda = 1$, giving $\tilde{\sigma} = \sigma$. □

Suppose that we receive $r(X)$ and wish to decode it.

- Compute $\sum_{j=1}^{2t} r(\alpha^j) X^j$.
- Set $\sigma(X) = 1 + \sigma_1 X + \dots + \sigma_t X^t$, and compute the coefficients of X^i for $t + 1 \leq i \leq 2t$ to obtain linear equations for $\sigma_1, \dots, \sigma_t$, which are of the form $\sum_0^t \sigma_j r(\alpha^{i-j}) = 0$.
- Then solve these polynomials over K, keeping solutions of least degree.
- Compute $\mathcal{E} = \left\{i \mid \sigma(\alpha^{-i}) = 0\right\}$, and check that $|\mathcal{E}| = \deg \sigma$.

- Set $e(X) = \sum_{i \in \mathcal{E}} X^i$, then $c(X) = r(X) + e(X)$, and check that c is a codeword.

Example. Consider $n = 7$, and $X^7 - 1 = (X+1)(X^3+X+1)(X^3+X^2+1)$ in $\mathbb{F}_2[X]$. Let $g(X) = X^3 + X + 1$, so $h(X) = (X+1)(X^3+X^2+1) = X^4 + X^2 + X + 1$. The parity check matrix is

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

The columns are the elements of $\mathbb{F}_2^3 \setminus \{0\}$. This is the Hamming $(7,4)$-code.

Let K be a splitting field for $X^7 - 1$; we can take $K = \mathbb{F}_8$. Let $\beta \in K$ be a root of g. Note that $\beta^3 = \beta + 1$, so $\beta^6 = \beta^2 + 1$, so $g(\beta^2) = 0$, and hence $g(\beta^4) = 0$. So the BCH code defined by $\{\beta, \beta^2\}$ has generator polynomial $g(X)$, again proving that this is Hamming's $(7,4)$-code. This code has design distance 3, so $d(C) \geq 3$, and we know Hamming's code has minimum distance exactly 3.

5.6. Shift registers

Definition. A *(general) feedback shift register* is a map $f : \mathbb{F}_2^d \to \mathbb{F}_2^d$ given by

$$f(x_0, \dots, x_{d-1}) = (x_1, \dots, x_{d-1}, C(x_0, \dots, x_{d-1}))$$

where $C : \mathbb{F}_2^d \to \mathbb{F}_2$. We say that the register has length d. The *stream* associated to an *initial fill* $(y_0, \dots, y_{d-1})$ is the sequence $y_0, \dots$ with $y_n = C(y_{n-d}, \dots, y_{n-1})$ for $n \geq d$.

Definition. The general feedback shift register $f : \mathbb{F}_2^d \to \mathbb{F}_2^d$ is a *linear feedback shift register* if C is linear, so

$$C(x_0, \dots, x_{d-1}) = \sum_{i=0}^{d-1} a_i x_i$$

We usually set $a_0 = 1$.

The stream produced by a linear feedback shift register is now given by the recurrence relation $y_n = \sum_{i=0}^{d-1} a_i y_{n-d+i}$. We can define the auxiliary polynomial $P(X) = X^d + a_{d-1}X^{d-1} + \dots + a_1 X + a_0$. We sometimes write $a_d = 1$, so $P(X) = \sum_{i=0}^{d} a_i X^i$.

Definition. The *feedback polynomial* is $\check{P}(X) = a_0 X^d + \dots + a_{d-1}X + 1 = \sum_{i=0}^{d} a_{d-i} X^i$. A sequence $y_0, \dots$ of elements of $\mathbb{F}_2$ has *generating function* $\sum_{j=0}^{\infty} y_j X^j \in \mathbb{F}_2[\![X]\!]$.

Theorem. The stream $(y_n)_{n \in \mathbb{N}}$ comes from a linear feedback shift register with auxiliary polynomial $P(X)$ if and only if its generating function is (formally) of the form $\frac{A(X)}{\check{P}(X)}$ with $A \in \mathbb{F}_2[X]$ such that $\deg A < \deg \check{P}$.

Note that $\check{P}(X) = X^{\deg P} P(X^{-1})$.

Proof. Let $P(X)$ and $\check{P}(X)$ be as above. We require

$$\left(\sum_{j=0}^{\infty} y_j X^j\right)\left(\sum_{i=0}^{d} a_{d-i} X^i\right)$$

to be a polynomial of degree strictly less than d. This holds if and only if the coefficient of X^n in $G(X)\check{P}(X)$ is zero for all $n \geq d$, which is $\sum_{i=0}^{d} a_{d-i} y_{n-i} = 0$. This holds if and only if $y_n = \sum_{i=0}^{d-1} a_i y_{n-d+i}$ for all $n \geq d$. This is precisely the form of a stream that arises from a linear feedback shift register with auxiliary polynomial P. □

The problem of recovering the linear feedback shift register from its stream and the problem of decoding BCH codes both involve writing a power series as a quotient of polynomials.

5.7. The Berlekamp–Massey method

Let $(x_n)_{n\in\mathbb{N}}$ be the output of a binary linear feedback shift register. We wish to find the unknown length d and values $a_0, \dots, a_{d-1}$ such that $x_n + \sum_{i=1}^{d} a_{d-i} x_{n-i} = 0$ for all $n \geq d$. We have

$$\underbrace{\begin{pmatrix} x_d & x_{d-1} & \cdots & x_1 & x_0 \\ x_{d+1} & x_d & \cdots & x_2 & x_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{2d-1} & x_{2d-2} & \cdots & x_d & x_{d-1} \\ x_{2d} & x_{2d-1} & \cdots & x_{d+1} & x_d \end{pmatrix}}_{A_d} \begin{pmatrix} a_d \\ a_{d-1} \\ \vdots \\ a_1 \\ a_0 \end{pmatrix} = 0$$

We look successively at $A_0 = \left(x_0\right), A_1 = \begin{pmatrix} x_1 & x_0 \\ x_2 & x_1 \end{pmatrix}, \dots$, starting at A_r if we know $d \geq r$. For each A_i, we compute its determinant. If $|A_i| \neq 0$, then $d \neq i$. If $|A_i| = 0$, we solve the system of linear equations on the assumption that $d = i$, giving a candidate for the coefficients $a_0, \dots, a_{d-1}$. This candidate can be checked over as many terms of the stream as desired.

6. Cryptography

6.1. Cryptosystems

We want to modify a message such that it becomes unintelligible to an eavesdropper Eve. Certain secret information is shared between two participants Alice and Bob, called the *key*, chosen from a set of possible keys $\mathcal{K}$. The unencrypted message is called the *plaintext*, which lies in a set $\mathcal{M}$, and the encrypted message is called the *ciphertext*, and lies in a set $\mathcal{C}$. A *cryptosystem* consists of $(\mathcal{K}, \mathcal{M}, \mathcal{C})$ together with the *encryption* function $e \colon \mathcal{M} \times \mathcal{K} \to \mathcal{C}$ and *decryption* function $d \colon \mathcal{C} \times \mathcal{K} \to \mathcal{M}$. These maps have the property that $d(e(m,k),k) = m$ for all $m \in \mathcal{M}, k \in \mathcal{K}$.

Example. Suppose $\mathcal{M} = \mathcal{C} = \{A, B, \dots, Z\}^\star = \Sigma^\star$. The *simple substitution cipher* defines $\mathcal{K}$ to be the set of permutations of Σ. To encrypt a message, each letter of plaintext is replaced with its image under a chosen permutation $\pi \in \mathcal{K}$.

The *Vigenère cipher* has $\mathcal{K} = \Sigma^d$ for some d. We identify Σ and $\mathbb{Z}/_{26\mathbb{Z}}$. Write out the key repeatedly below the plaintext, and add each plaintext letter with the corresponding key letter to produce a letter of ciphertext. For instance, encrypting the plaintext ATTACKATDAWN with the key LEMON gives ciphertext LXFOPVEFRNHR. Note, for instance, that each occurrence of the letter A in the plaintext corresponds to a letter of the key in the ciphertext. If $d = 1$, this is the *Caesar cipher*.

6.2. Breaking cryptosystems

Eve may know e and d, as well as the probability distributions of $\mathcal{K}, \mathcal{M}$, but she does not know the key itself. She seeks to recover the plaintext from a given string of ciphertext. There are three possible attack levels.

1. (ciphertext-only) Eve only knows some piece of ciphertext.
2. (known-plaintext) Eve knows a considerable length of plaintext and its corresponding ciphertext, but not the key. In other words, she knows m and $e(m,k)$, but not k.
3. (chosen plaintext) Eve can acquire the ciphertext for any plaintext message; she can generate $e(m,k)$ for any m.

Remark. The simple substitution cipher and Vigenère cipher fail at Level 1 in English if the messages are sufficiently long, as we can perform frequency analysis. Even if the plaintext is suitably random, both examples can fail at Level 2. For modern applications, Level 3 security is desirable.

Consider a cryptosystem $(\mathcal{M}, \mathcal{K}, \mathcal{C})$. We model the keys and messages as independent random variables K, M taking values in $\mathcal{K}, \mathcal{M}$. The ciphertext random variable is $C = e(K, M) \in \mathcal{C}$.

Definition. A cryptosystem $(\mathcal{M}, \mathcal{K}, \mathcal{C})$ has *perfect secrecy* if $H(M \mid C) = H(M)$, or equivalently, M and C are independent, or $I(M; C) = 0$.

One can show that perfect secrecy implies that $|\mathcal{K}| \geq |\mathcal{M}|$.

Definition. The *message equivocation* is $H(M \mid C)$. The *key equivocation* is $H(K \mid C)$.

Lemma. $H(M \mid C) \leq H(K \mid C)$.

Proof. Note that $M = d(C, K)$, hence $H(M \mid C, K) = 0$. Therefore, $H(C, K) = H(M, C, K)$. So

$$\begin{aligned} H(K \mid C) &= H(K, C) - H(C) \\ &= H(M, C, K) - H(M \mid K, C) - H(C) \\ &= H(M, K, C) - H(C) \\ &= H(K \mid M, C) + H(M, C) - H(C) \\ &= H(K \mid M, C) + H(M \mid C) \end{aligned}$$

Hence $H(K \mid C) \geq H(M \mid C)$. □

Let $\mathcal{M} = \mathcal{C} = \mathcal{A}$, and suppose we send n messages modelled as $M^{(n)} = (M_1, \dots, M_n)$ encrypted as $C^{(n)} = (C_1, \dots, C_n)$ using the same key K.

Definition. The *unicity distance* is the least n such that $H\big(K \mid C^{(n)}\big) = 0$; it is the smallest number of encrypted messages required to uniquely determine the key.

Now,

$$\begin{aligned} H\big(K \mid C^{(n)}\big) &= H\big(K, C^{(n)}\big) - H\big(C^{(n)}\big) \\ &= H\big(K, M^{(n)}, C^{(n)}\big) - H\big(C^{(n)}\big) \\ &= H\big(K, M^{(n)}\big) - H\big(C^{(n)}\big) \\ &= H(K) + H\big(M^{(n)}\big) - H\big(C^{(n)}\big) \end{aligned}$$

as $K, M^{(n)}$ are independent. We make the following assumptions.

(i) All keys are equally likely, so $H(K) = \log|\mathcal{K}|$.

(ii) $H\big(M^{(n)}\big) \approx nH$ for some constant H and sufficiently large n.

(iii) All sequences of ciphertext are equally likely, so $H\big(C^{(n)}\big) = n \log|\mathcal{A}|$.

Hence,

$$H\big(K \mid C^{(n)}\big) = \log|\mathcal{K}| + nH - n\log|\mathcal{A}|$$

This is nonnegative if and only if

$$n \leq U = \frac{\log|\mathcal{K}|}{\log|\mathcal{A}| - H}$$

Equivalently, $\frac{\log|\mathcal{A}|}{R\log|\mathcal{A}|}$ where $R = 1 - \frac{H}{\log|\mathcal{A}|}$ is the *redundancy* of the source. Recall that $0 \leq H \leq \log|\mathcal{A}|$. To make the unicity distance large, we can make the number of keys large, or use a message source with little redundancy.

6.3. One-time pad

Consider streams in $\mathbb{F}_2$ representing the plaintext $p_0, p_1, \dots$, the key stream $k_0, k_1, \dots$, and the ciphertext $z_0, z_1, \dots$ where $z_n = p_n + k_n$.

Definition. A *one-time pad* is a cryptosystem where k is generated randomly; the k_i are independent and take values of 0 or 1 with probability $\frac{1}{2}$.

$z = p + k$ is now a stream of independent and identically distributed random variables taking values of 0 or 1 with probability $\frac{1}{2}$. Hence, without the key stream, deciphering is impossible, so the unicity distance is infinite. One can show that a one-time pad has perfect secrecy.

In order to effectively use a one-time pad, we need to generate a random key stream. We then need to share the key stream to the recipient, which is exactly the initial problem. In most applications, the one-time pad is not practical. Instead, we share an initial fill $k_0, \dots, k_{d-1}$ to be used in a shared feedback shift register of length d to generate k. We then apply the following result.

Lemma. Let $x_0, x_1, \dots$ be a stream in $\mathbb{F}_2$ produced by a feedback shift register of length d. Then there exist $M, N \leq 2^d$ such that $x_{N+r} = x_r$ for all $r \geq M$.

Proof. Let the register be $f : \mathbb{F}_2^d \to \mathbb{F}_2^d$, and let $v_i = (x_i, \dots, x_{i+d-1})$. Then for all i, we have $f(v_i) = v_{i+1}$. Since $|\mathbb{F}_2^d| = 2^d$, the tuples $v_0, v_1, \dots, v_{2^d}$ cannot all be distinct. Let $a < b \leq 2^d$ such that $v_a = v_b$. Let $M = a$ and $N = b - a$, so $v_M = v_{M+N}$ so by induction we have $v_r = v_{r+N}$ for all $r \geq M$. □

Remark. The maximum period of a feedback shift register of length d is 2^d. For a linear feedback shift register, the maximum period is $2^d - 1$; this result is shown on the fourth example sheet.

Stream ciphers using linear feedback shift registers fail at level 2 due to the Berlekamp–Massey method. However, this cryptosystem is cheap, fast, and easy to use. Encryption and decryption can be performed on-the-fly, without needing the entire codeword first, and it is error-tolerant.

Recall that the stream produced by a linear feedback shift register is given by

$$x_n = \sum_{i=1}^{d} a_{d-i} x_{n-i}$$

for all $n \geq d$, and has auxiliary polynomial

$$P(X) = X^d + a_{d-1}X^{d-1} + \dots + a_0$$

with $a_d = 1$. The solutions to the recursion relations are linear combinations of powers of roots of P. Over $\mathbb{C}$, the general solution is a linear combination of $\alpha^n, n\alpha^n, \dots, n^{t-1}\alpha^n$ where α is a root of $P(X)$ with multiplicity t.

As $n^2 = n$ in $\mathbb{F}_2$, we cannot use this method directly. First, we must work in a splitting field K of P, a field containing $\mathbb{F}_2$ in which P is expressible as a product of linear factors. In addition, we replace the $n^i\alpha^n$ term with $\binom{n}{i}\alpha^n$. The general solution is now a linear combination of these terms in K.

We can also generate new key streams from old ones.

Lemma. Let (x_n), (y_n) be outputs from linear feedback shift registers of length M, N respectively. Then,

(i) the sequence $(x_n + y_n)$ is the output of a linear feedback shift register of length $M + N$;

(ii) the sequence $(x_n y_n)$ is the output of a linear feedback shift register of length MN.

The following proof is non-examinable.

Proof. Assume for simplicity that the auxiliary polynomials $P(X), Q(X)$ each have distinct roots α_1, α_M and $\beta_1, \dots, \beta_N$ in a field K extending $\mathbb{F}_2$. Then $x_n = \sum_{i=1}^{M} \lambda_i \alpha_i^n$ and $y_n = \sum_{i=1}^{N} \mu_j \beta_j^n$ where $\lambda_i, \mu_j \in K$. Now, $x_n + y_n = \sum_{i=1}^{M} \lambda_i \alpha_i^n \sum_{i=1}^{N} \mu_j \beta_j^n$ is produced by a linear feedback shift register with auxiliary polynomial $P(X)Q(X)$. For the second part, $x_n y_n = \sum_{i=1}^{M} \sum_{j=1}^{n} \lambda_i \mu_j (\alpha_i \beta_j)^n$ is the output of a linear feedback shift register with auxiliary polynomial $\prod_{i=1}^{N} \prod_{j=1}^{M} (X - \alpha_i \beta_j)$. □

Adding outputs of linear feedback shift registers is no more economical than producing the same string with a single linear feedback shift register. Muliplying streams does increase the effective length of the linear feedback shift register, but $x_n y_n = 0$ when either x_n or y_n are zero, so we gain little extra data. Nonlinear feedback shift registers are in general hard to analyse; in particular, an eavesdropper may understand the feedback shift register better than Alice and Bob.

6.4. Asymmetric ciphers

Stream ciphers are examples of symmetric cryptosystems. In such a system, the decryption process is the same, or is easily deduced from, the encryption process. In an asymmetric cryptosystem, the key is split into two parts: the *private key* for decryption, and the *public key* for encryption. Knowing the encryption and decryption processes and the public key, it should still be hard to find the private key or to decrypt the messages. This aim implies security at level 3. In this case, there is also no key exchange problem, since the public key can be broadcast on an open channel.

We base asymmetric cryptosystems on certain mathematical problems in number theory which are believed to be ‘hard’, such as the following.

(i) Factoring. Let $N = pq$ for p, q large prime numbers. Given N, the task is to find p and q.

(ii) Discrete logarithm problem. Let p be a large prime and g be a primitive root mod p (a generator of $\mathbb{F}_p^\star$). Given x, we wish to find a such that $x \equiv g^a$ mod p.

Definition. An algorithm runs in *polynomial time* if the number of operations needed to perform the algorithm is at most cN^d where N is the input size, and c, d are constants.

Example. An algorithm for factoring N has input size $\log_2 N$, roughly the number of bits in its binary expansion. Polynomial time algorithms include arithmetic operations on integers including the division algorithm, computation of greatest common divisors, and the Euclidean algorithm. We can also compute x^α mod N in polynomial time using repeated squaring; this is called modular exponentiation. Primality testing can be performed in polynomial time.

Polynomial time algorithms are not known for examples (i) and (ii) above. However, we have elementary methods for computing (i) and (ii) that take exponential time. If $N = pq$, dividing N by successive primes up to $\sqrt{N}$ will find p and q but takes $O(\sqrt{N}) = O(2^{\frac{B}{2}})$ steps where $B = \log_2 N$.

We describe the *baby-step, giant-step* algorithm for the discrete logarithm problem. Set $m = \lceil\sqrt{p}\rceil$, and write $a = qm + r$ for $0 \leq q, r < m$. Then, $x \equiv g^a = g^{qm+r}$ mod p, so $g^{qm} = g^{-r}x$ mod p. We list all values of g^{qm} and $g^{-r}x$ mod p; we then sort the lists and search for a match. This takes $O(\sqrt{p}\log p)$ steps.

The best known methods for solving the examples above use a factor base method, called the *modular number sieve.* It has running time

$$O\left(\exp\left(c(\log N)^{\frac{1}{3}}(\log\log N)^{\frac{2}{3}}\right)\right)$$

where c is a known constant.

6.5. Rabin cryptosystem

Recall that *Euler's totient function* is denoted φ, where $\varphi(n)$ is the number of integers less than n which are coprime to n. Equivalently, $\varphi(n) = \left|\left(\mathbb{Z}/n\mathbb{Z}\right)^\times\right|$. By Lagrange's theorem, $a^{\varphi(N)} \equiv 1$ mod N for each a coprime to N; this result is sometimes known as the Fermat–Euler theorem. If $N = p$ is prime, $a^{p-1} \equiv 1$ mod p, which is Fermat's little theorem.

Lemma. Let $p = 4k - 1$ be a prime, and let $d \in \mathbb{Z}$. If $x^2 \equiv d$ mod p is soluble, one solution is $x \equiv d^k$ mod p.

Proof. Suppose x_0 is a solution, so $x_0^2 \equiv d$ mod p. Without loss of generality we can assume $x_0 \not\equiv 0$, or equivalently, $x_0 \nmid p$. Then $x_0^2 \equiv d$ so $d^{2k-1} \equiv x_0^{2(2k-1)} \equiv x_0^{p-1} \equiv 1$. Hence, $\left(d^k\right)^2 \equiv d$. □

In the Rabin cryptosystem, the private key consists of two large distinct primes $p, q \equiv 3 \bmod 4$. The public key is $N = pq$. $\mathcal{M} = \mathcal{C} = \{1, \dots, N-1\} = \mathbb{Z}_N^\times$. We encrypt a plaintext message m as $c = m^2 \bmod N$. Usually, we restrict our messages so that $(m, N) = 1$ and $m > \sqrt{N}$.

Receiving ciphertext c, we can solve for x_1, x_2 such that $x_1^2 \equiv c \bmod p$ and $x_2^2 \equiv c \bmod q$ using the previous lemma. Then, applying the Chinese remainder theorem, we can find x such that $x \equiv x_1 \bmod p$ and $x \equiv x_2 \bmod q$, hence $x^2 \equiv c \bmod N$. Indeed, running the Euclidean algorithm on p, q gives integers r, s such that $rp + sq = 1$, then we can take $x = sqx_1 + rpx_2$.

Lemma. (i) Let p be an odd prime, and let $(d, p) = 1$. Then $x^2 \equiv d \bmod p$ has no solutions or exactly two solutions.

(ii) Let $N = pq$ where p, q are distinct odd primes, and let $(d, N) = 1$. Then $x^2 \equiv d \bmod N$ has no solutions or exactly four solutions.

Proof. Part (i). $x^2 \equiv y^2 \bmod p$ if and only if $p \mid (x^2 - y^2) = (x-y)(x+y)$, so either $p \mid x - y$ or $p \mid x + y$, so $x = \pm y$.

Part (ii). If x_0 is a solution, then by the Chinese remainder theorem, there exist solutions x with $x \equiv \pm x_0 \bmod p$ and $x \equiv \pm x_0 \bmod q$. This gives four solutions as required. By (i), these are the only possible solutions. □

Hence, to decrypt the Rabin cipher, we must find all four solutions to $x^2 \equiv c \bmod N$. Messages should include enough redundancy to uniquely determine which of these four solutions is the intended plaintext.

Theorem. Breaking the Rabin cryptosystem is essentially as difficult as factoring N.

Proof. If we can factorise N as pq, we have seen that we can decrypt messages. Conversely, suppose we can break the cryptosystem, so we have an algorithm to find square roots modulo N. Choose $x \bmod N$ at random, and use the algorithm to find y such that $y^2 \equiv x^2 \bmod N$. With probability $\frac{1}{2}$, $x \not\equiv \pm y \bmod N$. Then, $(N, x - y)$ is a nontrivial factor of N. If this fails, choose another x, and repeat until the probability of failure $\left(\frac{1}{2}\right)^r$ is acceptably low. □

6.6. RSA cryptosystem

Suppose $N = pq$ where p, q are distinct odd primes. We claim that if we know a multiple m of $\varphi(N) = (p-1)(q-1)$, then factoring N is 'easy'. Write $o_p(x)$ for the order of x as an element of $\left(\mathbb{Z}/p\mathbb{Z}\right)^\times$. Write $m = 2^a b$ where $a \geq 1, b$ odd. Let

$$X = \left\{ x \in \left(\mathbb{Z}/N\mathbb{Z}\right)^\times \,\middle|\, o_p(x^b) \neq o_q(x^b) \right\}$$

Theorem. (i) If $x \in X$, then there exists $0 \leq t < a$ such that $(x^{2^t b} - 1, N)$ is a nontrivial factor of N.

(ii) $|X| \geq \frac{1}{2}\left|\left(\mathbb{Z}/N\mathbb{Z}\right)^\times\right| = \frac{1}{2}(p-1)(q-1)$.

Proof. *Part (i).* By the Fermat–Euler theorem, $x^{\varphi(N)} \equiv 1 \bmod N$. Hence $x^m \equiv 1 \bmod N$. But $m = 2^a b$, so setting $y = x^b \bmod N$, we obtain $y^{2^a} \equiv 1 \bmod N$. In particular, $o_p(y)$ and $o_q(y)$ are powers of 2. Since $x \in X$, $o_p(y) \neq o_q(y)$, so without loss of generality suppose $o_p(y) < o_q(y)$. Let $o_p(y) = 2^t$, so $0 \leq t < a$. Then $y^{2^t} \equiv 1 \bmod p$, but $y^{2^t} \not\equiv 1 \bmod q$. So $(y^{2^t} - 1, N) = p$ as required. □

The proof of part (ii) will be seen later.

In the RSA cryptosystem, the private key consists of large distinct primes p, q chosen at random. Let $N = pq$, and choose the *encrypting exponent* e randomly such that $(e, \varphi(N)) = 1$, for instance taking e prime larger than p, q. By Euclid's algorithm, there exist d, k such that $de - k\varphi(N) = 1$; d is called the *decrypting exponent*.

The public key is (N, e), and we encrypt $m \in \mathcal{M}$ as $c \equiv m^e \bmod N$. The private key is (N, d), and we decrypt $c \in \mathcal{C}$ as $x \equiv c^d \bmod N$. By the Fermat–Euler theorem, $x \equiv m^{de} \equiv m^{1+k\varphi(N)} \equiv m \bmod N$, noting that the probability that $(m, N) \neq 1$ is small enough to be ignored. Hence, the decrypting function is inverse to the encrypting function.

Corollary. Finding the RSA private key (N, d) is essentially as difficult as factoring N.

Proof. We have already shown that if we can factorise N, we can find d. Conversely, suppose there is an algorithm to find d given N and e. Then $de \equiv 1 \bmod \varphi(N)$. Taking $m = de - 1$ in the proof of part (i) of the theorem above, we can factorise N. If this fails, repeat until the probability of failure is acceptably low. After r such random choices, we find a factor of N with probability $1 - \left(\frac{1}{2}\right)^r$. □

We now prove part (ii) of the above theorem.

Proof. The Chinese remainder theorem provides a multiplicative group isomorphism

$$\left(\mathbb{Z}/N\mathbb{Z}\right)^\times \to \left(\mathbb{Z}/p\mathbb{Z}\right)^\times \times \left(\mathbb{Z}/q\mathbb{Z}\right)^\times$$

mapping x to $(x \bmod p, x \bmod q)$. We claim that if we partition $\left(\mathbb{Z}/p\mathbb{Z}\right)^\times$ according to the value of $o_p(x^b)$, then each equivalence class has size at most

$$\frac{1}{2}\left|\left(\mathbb{Z}/p\mathbb{Z}\right)^\times\right| = \frac{1}{2}(p-1)$$

We show that one of these subsets has size exactly $\frac{1}{2}(p-1)$. Let g be a primitive root mod p, so $\left(\mathbb{Z}/p\mathbb{Z}\right)^\times = \langle g \rangle$. By Fermat's little theorem, $g^{p-1} \equiv 1 \bmod p$, so $g^m = g^{2^a b} \equiv 1 \bmod p$.

Hence, $o_p(g^b)$ is a power of 2, say $2^t \leq a$. Let $x = g^k$ for some $0 \leq k \leq p-2$, then $x^b = (g^b)^k$, so $o_p(x^b) = \frac{2^t}{(2^t,k)}$. So $o_p(x^b) = 2^t$ if and only if k is odd, so

$$o_p(x^b) = o_p(g^{bk}) = \begin{cases} o_p(g^b) = 2^t & \text{if } k \text{ odd} \\ < 2^t & \text{if } k \text{ even} \end{cases}$$

Thus, $\{g^k \bmod p \mid k \text{ odd}\}$ is the set as required, proving the claim. To finish, for each $y \in \left(\mathbb{Z}/q\mathbb{Z}\right)^\times$, the set

$$\left\{x \in \left(\mathbb{Z}/p\mathbb{Z}\right)^\times \,\middle|\, o_p(x^b) \neq o_q(x^b)\right\}$$

has at least $\frac{1}{2}(p-1)$ elements. Applying the Chinese remainder theorem,

$$|X| = \left|\left\{(x,y) \in \left(\mathbb{Z}/p\mathbb{Z}\right)^\times \times \left(\mathbb{Z}/q\mathbb{Z}\right)^\times \,\middle|\, o_p(x^b) \neq o_q(x^b)\right\}\right| \geq \frac{1}{2}(p-1)(q-1) = \frac{1}{2}\varphi(N)$$

□

Remark. We have shown that finding (N, d) from the public key (N, e) is as hard as factoring N. It is unknown whether decrypting messages sent via RSA is as hard as factoring.

RSA avoids the issue of needing to share keys, but it is slow. Symmetric ciphers are often faster.

Example (Shamir's padlock example). Let $\mathcal{A} = \mathbb{Z}_p$. Alice chooses $a \in \mathbb{Z}_{p-1}^\star$ and computes g^a. She finds a' such that $aa' = 1 \bmod p-1$. Bob chooses $b \in \mathbb{Z}_{p-1}^\star$ and computes g^b. He similarly finds b' such that $bb' = 1 \bmod p-1$.

Let m be a message in $\mathbb{Z}_p$. She encodes m as $c = m^a \bmod p$. She then sends this to Bob, who computes $d = c^b \bmod p$. He sends this back to Alice, who computes $e = d^{a'} \bmod p$. She sends this back to Bob, who computes $e^{b'} \bmod p$. By Fermat's little theorem, $e^{b'} \equiv d^{a'b'} \equiv c^{ba'b'} \equiv m^{aba'b'} \equiv m$.

$$m \xrightarrow{A} m^a \xrightarrow{B} c^b \xrightarrow{A} d^{a'} \xrightarrow{B} e^{b'}$$

Example (Diffie–Hellman key exchange). Alice and Bob wish to agree on a secret key k. Let p be a large prime, and g a primitive root mod p. Alice chooses an exponent $\alpha \in \mathbb{Z}_{p-1}$ and sends $g^\alpha \bmod p$ to Bob. Bob chooses an exponent β and sends $g^\beta \bmod p$ to Alice. Both Alice and Bob compute $k = g^{\alpha\beta}$, which can be used as their secret key. An eavesdropper must find $g^{\alpha\beta}$ knowing g, g^α, and g^β. Diffie and Hellman conjectured that this problem is as difficult as solving the discrete logarithm problem.

6.7. Secrecy and attacks

Consider a message m sent by Alice to Bob. Here are some possible aims that the participants may have in communication.

(i) *Secrecy*: Alice and Bob can be sure that no third party can read the message.

(ii) *Integrity*: Alice and Bob can be sure that no third party can alter the message.

(iii) *Authenticity*: Bob can be sure that Alice sent the message.

(iv) *Non-repudiation*: Bob can prove to a third party that Alice sent the message.

Example (authenticity using RSA)**.** Suppose Alice uses a private key (N, d) to encrypt m. Anyone can decrypt m using the public key (N, e) as $(m^d)^e = (m^e)^d = m$, but they cannot forge a message sent by Alice. Suppose Bob picks a random message m and sends it to Alice; if Bob then receives a message back from Alice which after decryption ends in m, then he can be sure it comes from Alice.

Signature schemes preserve integrity and non-repudiation. They also prevent tampering in the following sense.

Example (homomorphism attack)**.** Suppose a bank sends messages of the form (M_1, M_2) where M_1 represents the client's name and M_2 represents an amount of money to be transferred into their account. Suppose that messages are encoded using RSA as $(Z_1, Z_2) = (M_1^e, M_2^e)$, where all calculations are performed modulo N. A client C transfers £100 to their account, and observes the encrypted message (Z_1, Z_2). Then, sending (Z_1, Z_2^3) to the bank, C becomes a millionaire without breaking RSA. Alternatively, one could simply send (Z_1, Z_2) to the bank many times, gaining more money each time; this particular attack is defeated by timestamping the messages.

Definition. A message m is *signed* as (m, s) where the *signature* $s = s(m, k)$ is a function of m and the private key k.

The recipient can check the signature using the public key to verify authenticity of the message. The signature function or *trapdoor* function $s\colon \mathcal{M} \times \mathcal{K} \to \mathcal{S}$ is designed such that without knowledge of the private key, one cannot sign messages, but anyone can check whether a signature is valid. Note that the signature is associated to each message, not to each sender.

Example (signatures using RSA)**.** Suppose Alice has a private key (N, d), and broadcasts a public key (N, e). She signs a message m as (m, s) where $s = m^d \bmod N$. The signature is verified by checking $s^e = m$.

This technique is vulnerable to the homomorphism attack. This is also vulnerable to the *existential forgery* attack, in which an attacker produces valid signed messages of the form $(s^e \bmod N, s)$ after choosing s first. Hopefully, such messages are not meaningful.

To solve these problems, we could use a better signature scheme. In addition, rather than signing a message m, we instead sign the *digest* $h(m)$ where $h\colon \mathcal{M} \to \{1, \dots, N-1\}$ is a *hash* function. A hash function is a publicly known function for which it is very difficult to find pairs of messages with matching hashes; such a pair is called a *collision*. Examples of hash functions include MD5 and the SHA family.

6.8. Elgamal signature scheme

Alice chooses a large prime p and a random integer u with $1 < u < p$. Let g be a primitive root mod p. The public key is $p, g, y = g^u \bmod p$. The private key is u. Let $h : \mathcal{M} \to \{1, \dots, p-1\}$ be a collision-resistant hash function.

To send a message m with $0 \leq m \leq p-1$, Alice randomly chooses k with $1 \leq k \leq p-2$ coprime to $p-1$. She computes r, s with $1 \leq r \leq p-1$ and $1 \leq s \leq p-2$ satisfying

$$r \equiv g^k \mod p; \quad h(m) \equiv ur + ks \mod (p-1)$$

Since k is coprime to $p-1$, the congruence for s always has a solution. Alice signs the message with the signature (r, s). Now,

$$g^{h(m)} \equiv g^{ur+ks} \equiv (g^u)^r (g^k)^s \equiv y^r r^s \mod p$$

Bob accepts a signature if $g^{h(m)} \equiv y^r r^s \bmod p$. To forge a signature, obvious attacks involve the discrete logarithm problem, finding u from $y = g^u$.

Lemma. Let $a, b, m \in \mathbb{N}$ and consider the congruence $ax \equiv b \bmod m$. This has either no solutions or $\gcd(a, m)$ solutions for $x \bmod m$.

Proof. Let $d = \gcd(a, m)$. If $d \nmid b$, there is no solution. If $d \mid b$, we can rewrite the congruence as $\frac{a}{d}x \equiv \frac{b}{d} \bmod \frac{m}{d}$. Note that $\frac{a}{d}, \frac{m}{d}$ are coprime, so this congruence has a unique solution. □

It is vital that Alice chooses a new value of k to sign each message. Suppose she sends m_1, m_2 using the same value of k. Denote the signatures (r, s_1) and (r, s_2); note that r depends only on k and is hence fixed.

$$h(m_1) \equiv ur + ks_1 \mod (p-1); \quad h(m_2) \equiv ur + ks_2 \mod (p-1)$$

Hence,

$$h(m_1) - h(m_2) \equiv k(s_1 - s_2) \mod (p-1)$$

Let $d = \gcd(p-1, s_1 - s_2)$. By the previous lemma, this is the number of solutions for k modulo $p-1$. Choose the solution that gives the correct value in the first congruence $r \equiv g^k \bmod p$. Then,

$$s_1 \equiv \frac{h(m_1) - ur}{k} \mod (p-1)$$

This gives $ur \equiv h(m_1) - ks_1$. Hence, using the lemma again, there are $\gcd(p-1, r)$ solutions for u. Choose the solution for u that gives $y \equiv g^u$. This allows us to deduce Alice's private key u, as well as the exponent k used in both messages.

6.9. The digital signature algorithm

The digital signature algorithm is a variant of the Elgamal signature scheme developed by the NSA. The public key is (p, q, g) constructed as follows.

- Let p be a prime of exactly N bits, where N is a multiple of 64 such that $512 \leq N \leq 1024$, so $2^{N-1} < p < 2^N$.
- Let q be a prime of 160 bits, such that $q \mid p - 1$.
- Let $g \equiv h^{\frac{p-1}{q}} \bmod p$, where h is a primitive root mod p; in particular, g is an element of order q in $\mathbb{Z}_p^\times$.
- Alice chooses a private key x with $1 < x < q$ and publishes $y = g^x$.

Let m be a message with $0 \leq m < q$. She chooses a random k with $1 < k < q$, and computes

$$s_1 \equiv (g^k \bmod p) \bmod q; \quad s_2 \equiv k^{-1}(m + xs_1) \bmod q$$

The signature is (s_1, s_2). To verify a signature, we perform the following procedure. Bob computes $w \equiv s_2^{-1} \bmod q$, $u_1 \equiv mw \bmod q$, $u_2 \equiv s_1 w \bmod q$, and $v = (g^{u_1} y^{u_2} \bmod p) \bmod q$. He accepts the signature if $v = s_1$.

Proposition. If a message is signed with the DSA and the message is not manipulated, the signature is accepted.

Proof. First, note that $(m + xs_1)w = ks_2 s_2^{-1} \bmod q$. Now, as $g^q = 1 \bmod p$,

$$\begin{aligned} v &= (g^{u_1} y^{u_2} \bmod p) \bmod q \\ &= (g^{mw} g^{xs_1 w} \bmod p) \bmod q \\ &= (g^{(m+xs_1)w} \bmod p) \bmod q \\ &= (g^k \bmod p) \bmod q \\ &= s_1 \end{aligned}$$

Hence, for a correctly signed message, the verification succeeds. □

Suppose that Alice sends m_1 to Bob and m_2 to Carol, and provides signatures for each message using the DSA. One can show that if Alice uses the same value of k for both transmissions, it is possible for an eavesdropper to recover the private key x from the signed messages.

6.10. Commitment schemes

Suppose Alice wants to send a bit $m \in \{0, 1\}$ to Bob in such a way that

(i) Bob cannot determine the value of m without Alice's help; and

(ii) Alice cannot change the bit once she has sent it.

Such a system can be used for coin tossing: suppose Alice and Bob are in different rooms, where Alice tosses a coin and Bob guesses the result. The result of the coin and Bob's guess can be viewed as messages of this form. As another example, consider a poll whose result cannot be viewed until everyone has voted. We will see two examples of such a *commit-and-reveal* strategy, known as *bit commitment*.

Suppose that we have a publicly known encryption function e_A and a decryption function d_A known only to Alice. Alice makes a choice for her message m, and commits to Bob the ciphertext $c = e_A(m)$. Under the assumption that the cipher is secure, Bob cannot decipher the message. To reveal her choice, Alice sends her private key to Bob, who can then use it to decipher the message $d_A(c) = d_A(e_A(m)) = m$. He can also check that d_A, e_A are inverse functions and thus ensure that Alice sent the correct private key.

Alternatively, suppose that Alice has two ways to communicate to Bob: a clear channel which transmits with no errors, and a binary symmetric channel with error probability p. Suppose $0 < p < \frac{1}{2}$, and the noisy channel corrupts bits independent of any action of Alice or Bob, so neither can affect its behaviour. Bob publishes a binary linear code C of length N and minimum distance d, and Alice publishes a random non-trivial linear map $\theta : C \to \mathbb{F}_2$. To send a bit $m \in \mathbb{F}_2$, Alice chooses a random codeword $c \in C$ such that $\theta(c) = m$, and sends c to Bob via the noisy channel. Bob receives $r = c + e \in \mathbb{F}_2^N$ where e is the error pattern. The expected value of $d(r, c) = d(e, 0)$ is Np. N is chosen such that $Np \gg d$, so Bob cannot tell what the original codeword c was, and hence cannot find $\theta(c) = m$.

To reveal, Alice sends c to Bob using the clear channel. Bob can check that $d(c, r) \approx Np$; if so, he accepts the message. It is possible that many more or many fewer bits of c were corrupted by the noisy channel, which may make Bob reject the message even if Alice correctly committed and revealed the message. N, d should be chosen such that the probability of this occurring is negligible.

We have shown that Bob cannot read Alice's guess until she reveals it. In addition, Alice cannot cheat by changing her guess, because she knows c but not how it was corrupted by the noisy channel. All she knows is that the received message r has distance approximately Np from c. If she were to send $c' \neq c$, she must ensure that $d(r, c') \approx Np$, but the probability that this happens is small unless she chooses c' very close to c. But any two distinct codewords have distance at least d, so she cannot cheat.

6.11. Secret sharing schemes

Suppose that the CMS is attacked by the MIO. The Faculty will retreat to a bunker known as MR2. Entry to MR2 is controlled by a *secret*, which is a positive integer S. This secret is known only to the Leader. Each of the n members of the Faculty knows a pair of numbers, called their *shadow* or *share*. It is required that, in the absence of the Leader, any k members of the Faculty can reconstruct the secret from their shadows, but any $k - 1$ cannot.

Definition. Let $k, n \in \mathbb{N}$ with $k < n$. A (k, n)*-threshold scheme* is a method of sharing a message S among a set of n participants such that any subset of k participants can reconstruct S, but no subset of smaller size can reconstruct S.

We discuss Shamir's method for implementing such a scheme. Let $0 \leq S \leq N$ be the secret, which can be chosen at random by the Leader. The Leader chooses and publishes a prime $p > n, N$. They then choose independent random coefficients $a_1, \dots, a_{k-1}$ with $0 \leq a_j \leq p-1$ where we take $a_0 = S$, and distinct integers $x_1, \dots, x_n$ with $1 \leq x_j \leq p-1$. Define

$$P(r) \equiv a_0 + \sum_{j=1}^{k-1} a_j x_r^j \mod p$$

choosing $0 \leq P(r) \leq p-1$. The rth participant is given their shadow pair $(x_r, P(r))$ to be kept secret. The Leader can then discard their computations.

Suppose k members of the Faculty assemble with shadow pairs $(y_j, Q(j)) = (x_{i_j}, P(i_j))$ for $1 \leq j \leq k$. By properties of the Vandermonde determinant,

$$\det \begin{pmatrix} 1 & y_1 & \cdots & y_1^{k-1} \\ 1 & y_2 & \cdots & y_2^{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & y_k & \cdots & y_k^{k-1} \end{pmatrix} = \prod_{1 \leq j < i \leq k} (y_i - y_j)$$

The y_i are distinct, so this determinant does not vanish. Hence, we can uniquely solve the system of k simultaneous equations

$$\begin{aligned} z_0 + y_1 z_1 + y_1^2 z_2 + \dots + y_k^{k-1} z_{k-1} &\equiv Q(1) \\ z_0 + y_2 z_1 + y_2^2 z_2 + \dots + y_2^{k-1} z_{k-1} &\equiv Q(2) \\ &\vdots \\ z_0 + y_k z_1 + y_k^2 z_2 + \dots + y_k^{k-1} z_{k-1} &\equiv Q(k) \end{aligned}$$

In particular, $z_0 = a_0 = S$ is the secret, as $(a_0, \dots, a_{k-1})$ is also a solution to these equations by construction. Suppose $k-1$ people attempt to reconstruct the secret. In this case, the Vandermonde determinant gives

$$\det \begin{pmatrix} y_1 & y_1^2 & \cdots & y_1^{k-1} \\ y_2 & y_2^2 & \cdots & y_2^{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k-1} & y_{k-1}^2 & \cdots & y_{k-1}^{k-1} \end{pmatrix} = y_1 y_2 \dots y_{k-1} \prod_{1 \leq j < i \leq k-1} (y_i - y_j)$$

This is nonzero modulo p, so the system of equations

$$\begin{aligned} z_0 + y_1 z_1 + y_1^2 z_2 + \dots + y_k^{k-1} z_{k-1} &\equiv Q(1) \\ z_0 + y_2 z_1 + y_2^2 z_2 + \dots + y_2^{k-1} z_{k-1} &\equiv Q(2) \\ &\vdots \\ z_0 + y_{k-1} z_1 + y_{k-1}^2 z_2 + \dots + y_{k-1}^{k-1} z_{k-1} &\equiv Q(k-1) \end{aligned}$$

has solutions for $z_1, \dots, z_{k-1}$ regardless of the value of z_0. Thus, $k-1$ members of the Faculty cannot reconstruct the secret S, or even tell which values are more likely than others.

Remark. Note that a polynomial of degree $k-1$ can be recovered from its values at k points, but not on fewer points; this technique is known as Lagrange interpolation. The secret shadow pairs can be changed without altering the secret S; the Leader simply chooses a different random polynomial with the same constant term. Changing the polynomial frequently can increase security, since any eavesdropper who has gathered some shadow pairs generated from one polynomial cannot use that information to help decrypt a different polynomial.

Example. Consider a $(3,n)$-threshold scheme, where ordinary workers in a company have single shares, the vice presidents have two shares, and the Leader has three. In this case, the secret can be recovered by any three ordinary workers, any two workers if one of them is a vice president, or the Leader alone. In such *hierarchical schemes*, the 'importance' of individuals determines how many of them are required to recover the secret.

Example. Suppose Alice has a private key that she wishes to store securely and reliably. She uses a $(k,2k-1)$-threshold scheme, where she forms $2k-1$ shadow pairs and stores them in different locations. As long as she does not lose more than half of the pairs, she can recover her key, hence the scheme is reliable. An eavesdropper needs to steal more than half of the pairs in order to recover the key, hence the scheme is secure.

VII. Quantum Information and Computation

Lectured in Lent 2023 by Prof. N. Datta

Computers manipulate bits of information to process inputs and answer questions. Regardless of the physical form of the computer, it has certain fundamental theoretical limitations. As the size of a bit string increases, the number of possible values of the string increases exponentially, and this means that many computational tasks require exponential amounts of time or space to compute a result.

Quantum computation allows us to bypass some of these limitations by leveraging features of quantum mechanics. In the quantum case, we store information using quantum bits ('qubits') instead of classical bits. While a classical computer can only operate on a single state at a time, we can construct a superposition of quantum states and operate on them all at once. We can use this to solve certain classical problems with a quantum computer faster than is possible with a classical computer, even in theory.

One example of a difficult problem is SAT, the Boolean satisfiability problem. The input is a Boolean function in n variables, and we wish to determine whether there is an assignment of the variables that makes the formula true. This problem is NP-complete: any problem in the complexity class NP can be reduced to a case of SAT. One of the quantum algorithms discussed in this course is Grover's quantum search algorithm, which solves SAT with a quadratic speedup compared to the classical complexity. This shows that Grover's algorithm can be applied to any NP problem to give a quadratic speedup. Hence, quantum computers can be used to solve a wide class of problems faster than a classical computer can.

Contents

1. Mathematical background

1.1. Motivation

In classical computation, the elementary unit of information is the *bit*, which takes a value in $\{0, 1\}$. This gives the result of a single binary decision problem, where the zero and one correspond to different answers to the problem. Binary strings of length greater than one are used to provide more than 2 answers to a problem; if we have n bits, we can encode 2^n different messages.

Classical computation is understood to be the processing of information: taking an initial bit string and and updating it by a prescribed sequence of steps. The steps are taken to be the action of local Boolean logic gates, such as conjunction, disjunction, or negation. At each step, a small number of bits in prescribed locations are edited.

Information in the real world must be tied to a physical representation. For example, bits in a processor are often represented by different voltages of specific components. Importantly, there is no information *without* representation. Performing a computation classically must therefore involve the evolution of a physical system over time, which is coverned by the laws of classical physics.

However, nature does not abide by classical physics at subatomic levels, and we must use quantum mechanics to accurately model such behaviours. One such behaviour modelled by quantum mechanics is the superposition principle, that the corresponding quantum analog of the bit need not be in precisely one state. Quantum entanglement is the phenomenon where particles can be linked in such a way that their states can be manipulated even at a distance. Quantum measurement is probabilistic and alters the underlying system.

Quantum information and computation therefore exploits these features of quantum mechanics to address issues of information storage, communication, computation, and cryptography. The features of quantum mechanics seem to allow us benefits which are beyond the limits of classical information and computation, even in principle. Note that a quantum computer cannot perform any task that cannot in principle be performed classically. We only hope that quantum techniques allow a reduction in the complexity of certain algorithms.

1.2. Benefits of quantum information and computation

In complexity theory, we study the *hardness* of a certain computational task. One must consider the resources required for the task; which in classical computation are normally limited to time (measured in number of computational steps) and space (amount of memory required).

If an algorithm takes time bounded by a polynomial function in the input size n, we say the algorithm is *polynomial-time*. Otherwise, we say it is an *exponential-time* algorithm. Polynomial-time algorithms are typically taken to be computable in practice, but exponential-time algorithms are usually considered only computable in principle. Quantum mechanical techniques can provide polynomial-time algorithms that have only exponential-time classical versions. One example is Shor's integer factorisation algorithm.

Quantum states of physical sytems can be used to encode information, such as spin states of electrons. There are certain tasks possible with such quantum states which are impossible in classical physics; one example is quantum teleportation.

There are also some technological issues with classical physics. Components of processors have become minified to atomic scale, and therefore they cannot be shrunk much further without dealing with the effects of quantum mechanics. Conversely, there are technological challenges with quantum physics. Quantum systems are very fragile, and modern quantum computers typically require temperatures close to absolute zero to reduce noise.

Quantum supremacy refers to the hypothetical moment at which a programmable quantum computer can first solve a problem in practice that a classical computer cannot. At the time of writing, there is no concensus that quantum supremacy has been achieved.

1.3. Hilbert spaces

Every quantum mechanical system is associated with a Hilbert space $\mathcal{V}$, a complex inner product space that is a complete metric space with respect to the distance function induced by the inner product. We use Dirac's *bra-ket* notation: a vector is represented by $|v\rangle \in \mathcal{V}$, and its conjugate transpose is denoted $\langle v| \in \mathcal{V}^\star$. If $\mathcal{V} = \mathbb{C}^n$, we write

$$|\psi\rangle = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}; \quad \langle\psi| = \begin{pmatrix} a_1^\star & \cdots & a_n^\star \end{pmatrix}$$

The inner product of ψ and ϕ is written $\langle\psi|\phi\rangle$. Recall that an inner product satisfies

- $\langle\psi|\psi\rangle \geq 0$, and equal to zero if and only if $|\psi\rangle = 0$;
- linearity in the second argument, so $\langle\psi|a\phi_1 + b\phi_2\rangle = a\langle\psi|\phi_1\rangle + b\langle\psi|\phi_2\rangle$;
- antilinearity in the first argument, so $\langle a\psi_1 + b\psi_2|\phi\rangle = a^\star\langle\psi_1|\phi\rangle + b^\star\langle\psi_2|\phi\rangle$;
- skew-symmetry, so $\langle\psi|\phi\rangle^\star = \langle\phi|\psi\rangle$;

and induces a norm $\|\psi\| = \||\psi\rangle\| = \sqrt{\langle\psi|\psi\rangle}$. In this course, we will often consider $\mathcal{V} = \mathbb{C}^2$ and define

$$|0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}; \quad |1\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

For an arbitrary $|v\rangle \in \mathbb{C}^2$, we can write $|v\rangle = a|0\rangle + b|1\rangle$, giving

$$|v\rangle = \begin{pmatrix} a \\ b \end{pmatrix}; \quad \langle v| = \begin{pmatrix} a^\star & b^\star \end{pmatrix}$$

If $|w\rangle = c|0\rangle + d|1\rangle$, then $\langle v|w\rangle = a^\star c + b^\star d$.

We can also compute the *outer product* of two vectors, defined to be $|\psi\rangle\langle\phi|$. If $\mathcal{V} = \mathbb{C}^n$, the outer product is an $n \times n$ matrix. An orthonormal basis $(|i\rangle)_{i=1}^n$ for $\mathcal{V}$ is called *complete* if $\sum_{i=1}^n |i\rangle\langle i|$ is the identity matrix.

If $\mathcal{V}$ has a complete orthonormal basis, we can write $|\psi\rangle = \sum_{i=1}^n c_i\,|i\rangle$ for some c_i. If $\langle\psi|\psi\rangle = 1$, we say $|\psi\rangle$ is *normalised*. In this case, $\sum |c_i|^2 = 1$, and the $|c_i|^2$ form a discrete probability distribution. We call the c_i the *probability amplitudes*.

Let $\mathcal{V}, \mathcal{W}$ be vector spaces, where $\dim \mathcal{V} = n, \dim \mathcal{W} = m$. Let $|v\rangle \in \mathcal{V}, |w\rangle \in \mathcal{W}$. Suppose $|v\rangle = \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix}^\intercal$, and $|w\rangle = \begin{pmatrix} b_1 & \cdots & b_m \end{pmatrix}^\intercal$. Then, $|v\rangle \otimes |w\rangle$ is the *tensor product* of $|v\rangle$ and $|w\rangle$, defined by

$$|v\rangle \otimes |w\rangle = \begin{pmatrix} a_1b_1 \\ \vdots \\ a_1b_m \\ a_2b_1 \\ \vdots \\ a_nb_m \end{pmatrix} \in \mathcal{V} \otimes \mathcal{W}$$

If $(|e_i\rangle)_{i=1}^n$ is a complete orthonormal basis for $\mathcal{V}$ and $\left(|f_j\rangle\right)_{j=1}^m$ is a complete orthonormal basis for $\mathcal{W}$, then $\left(|e_i\rangle \otimes |f_j\rangle\right)_{i,j=1}^{n,m}$ is a complete orthonormal basis for $\mathcal{V}\otimes\mathcal{W}$. We sometimes write $|v\rangle \otimes |w\rangle$ as $|v\rangle\,|w\rangle$ or $|vw\rangle$.

If $|\alpha\rangle \in \mathcal{V}$, we can write $|\alpha\rangle = \sum a_i\,|e_i\rangle$, and similarly if $|\beta\rangle \in \mathcal{W}$, we can write $|\beta\rangle = \sum b_j\,|f_j\rangle$. Then, $|\alpha\beta\rangle = \sum a_i c_j\,|e_i f_j\rangle$.

We say $|\Psi\rangle \in \mathcal{V} \otimes \mathcal{W}$ is a *product vector* if $|\Psi\rangle = |\psi\rangle \otimes |\phi\rangle$ for some ψ, ϕ. Vectors that are not product vectors are called *entangled vectors*.

Let $\mathcal{V} = \mathbb{C}^2 = \mathcal{W}$. Define $|\phi^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$. Suppose $|\phi^+\rangle = |\psi\rangle \otimes |\phi\rangle = (a\,|0\rangle + b\,|1\rangle) \otimes (c\,|0\rangle + d\,|1\rangle)$. Then, $|\phi^+\rangle = ac\,|00\rangle + ad\,|01\rangle + bc\,|10\rangle + bd\,|11\rangle$. So one of a and d, and one of b and c is equal to zero, contradicting the assumption, so $|\phi^+\rangle$ is entangled.

We define the inner product on the product space by defining

$$\langle\phi_1|\psi_2\rangle = (\langle\alpha_1|\,\langle\beta_1|)(|\beta_2\rangle\,|\alpha_2\rangle) = \langle\alpha_1|\alpha_2\rangle\,\langle\beta_1|\beta_2\rangle$$

where $|\psi_i\rangle = |\alpha_i\rangle\,|\beta_i\rangle$. In the general case, $|A\rangle = \sum a_{ij}\,|e_i\rangle\,|f_j\rangle, |B\rangle = \sum b_{ij}\,|e_i\rangle\,|f_j\rangle$, and we define

$$\langle A|B\rangle = \left(\sum a_{ij}^\star\,\langle e_i|\,\langle f_j|\right)\left(\sum b_{ij}\,|e_i\rangle\,|f_j\rangle\right) = \sum a_{ij}^\star b_{ij}\delta_{ii'}\delta_{jj'} = \sum a_{ij}^\star b_{ij}$$

where δ is the Kronecker δ symbol.

We define the k-fold *tensor power* of a vector space $\mathcal{V}$ by

$$\mathcal{V}^{\otimes n} = \underbrace{\mathcal{V} \otimes \cdots \otimes \mathcal{V}}_{n \text{ times}}$$

If $\mathcal{V} = \mathbb{C}^2$, this has dimension 2^k, and complete orthonormal basis $|i_1 \dots i_k\rangle$ for $i_j \in \{0, 1\}$. Note that $|v\rangle\,|w\rangle \neq |w\rangle\,|v\rangle$.

1.4. First postulate: quantum states

In this course, we will restrict our attention to finite-dimensional vector spaces, and finite time evolution. We describe the *postulates* for quantum mechanics that we will work under.

The first postulate is that, given an isolated quantum mechanical system S, we can associate a finite-dimensional vector space $\mathcal{V}$. The physical state of the system is given by a unit vector $|\psi\rangle$ in $\mathcal{V}$. More precisely, the state is given by a *ray*, an equivalence class of vectors $e^{i\theta}|\psi\rangle$ for $\theta \in \mathbb{R}$. No measurements can distinguish states in a given equivalence class. Note that states $a|\psi_1\rangle + b|\psi_2\rangle$ and $a|\psi_1\rangle + be^{i\theta}|\psi_2\rangle$ can be distinguished by measurement, since the phase difference is relative, not global.

Example. Let $\mathcal{V} = \mathbb{C}^2$ with (complete orthonormal) basis $|0\rangle, |1\rangle$. The elementary unit of quantum information is known as the *qubit*, which is any quantum system with $\mathcal{V} = \mathbb{C}^2$. The spin of an electron, which is some superposition of spin-up and spin-down, can be modelled by $\mathbb{C}^2$. A property of the polarisation of a photon, such as vertical or horizontal, or right-circular or left-circular, can also be modelled in this way.

Define $|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$ and $|-\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$. This is another complete orthonormal basis for $\mathcal{V}$, sometimes called the *conjugate basis*.

1.5. Second postulate: composite systems

The second postulate of quantum mechanics is that two quantum systems S_1, S_2 with associated vector spaces $\mathcal{V}_1, \mathcal{V}_2$ can be composed into the *composite system* with vector space $\mathcal{V}_1 \otimes \mathcal{V}_2$.

Example. Consider $\mathcal{V}^{\otimes n}$, the space of n qubits. An orthonormal basis is $|i_1 \dots i_n\rangle$ where $i_j \in \{0, 1\}$. A vector in $\mathcal{V}^{\otimes n}$ can be written $\sum a_{i_1 \dots i_n} |i_1 \dots i_n\rangle$. There are 2^n different amplitudes $a_{i_1 \dots i_n}$, providing exponential growth in information. However, in a product state, we obtain only linear growth in information.

1.6. Observables

An *observable* is a property of a physical system which can, in theory, be measured. Mathematically, these are modelled by linear self-adjoint (or Hermitian) operators.

The action of a linear operator A on a state space $\mathcal{V}$ is a written $A|\psi\rangle$. By linearity, we have $A(a|\psi\rangle + b|\phi\rangle) = aA|\psi\rangle + bA|\phi\rangle$ for $a, b \in \mathbb{C}$. For any operator A acting on $\mathcal{V}$, there is a unique linear operator $A^\dagger$ such that $\langle v|Aw\rangle = \langle A^\dagger v|w\rangle$, called the *adjoint* of A; operators equal to their adjoints are called *self-adjoint*.

We can easily show that $(AB)^\dagger = B^\dagger A^\dagger$. By convention, we define $|\psi\rangle^\dagger = \langle\psi|$, so for a self-adjoint operator A, we have $(A|\psi\rangle)^\dagger = \langle\psi|A$. There are four important operators which act on the single-qubit space $\mathbb{C}^2$.

$$\sigma_0 = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix};\quad \sigma_x = \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix};\quad \sigma_y = \begin{pmatrix}0 & -i\\ i & 0\end{pmatrix};\quad \sigma_z = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}$$

σ_0 is the identity matrix, and $\sigma_x, \sigma_y, \sigma_z$ are called the *Pauli matrices*. The actions of these matrices on the basis vectors $|0\rangle$ and $|1\rangle$ are

$$\sigma_0 |0\rangle = |0\rangle;\quad \sigma_0 |1\rangle = |1\rangle;\quad \sigma_x |0\rangle = |1\rangle;\quad \sigma_x |1\rangle = |0\rangle;$$

$$\sigma_y |0\rangle = i\,|1\rangle;\quad \sigma_y |1\rangle = -i\,|0\rangle;\quad \sigma_z |0\rangle = |0\rangle;\quad \sigma_z |1\rangle = -|1\rangle$$

Note that

$$\sigma_x\sigma_y = i\sigma_z;\quad \sigma_y\sigma_z = i\sigma_x;\quad \sigma_z\sigma_x = i\sigma_y$$

Intuitively, σ_x is a bit flip, σ_y is a phase flip, and σ_z is a combined bit and phase flip.

1.7. Dirac notation for linear operators

Let $|v\rangle = a\,|0\rangle + b\,|1\rangle$, and $|w\rangle = c\,|0\rangle + d\,|1\rangle$. The outer product is

$$M = |v\rangle\langle w| = \begin{pmatrix}a\\ b\end{pmatrix}\begin{pmatrix}c^\star & d^\star\end{pmatrix} = \begin{pmatrix}ac^\star & ad^\star\\ bc^\star & bd^\star\end{pmatrix}$$

which is a linear map on $\mathcal{V} = \mathbb{C}^2$. One can show that $M\,|x\rangle = (|v\rangle\langle w|)\,|x\rangle = |v\rangle\,\langle w|x\rangle$, which is the scalar product of the vector $|v\rangle$ with the inner product $\langle w|x\rangle$. Such outer products yield the linear maps from $\mathbb{C}^2$ to $\mathbb{C}^2$ that have rank 1, and the kernel of M is the subspace of vectors orthogonal to $|w\rangle$. Note that

$$|0\rangle\langle 0| = \begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix};\quad |0\rangle\langle 1| = \begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix};\quad |1\rangle\langle 0| = \begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix};\quad |1\rangle\langle 1| = \begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix}$$

Hence, we can write

$$A = \begin{pmatrix}a & b\\ c & d\end{pmatrix} \implies A = a\,|0\rangle\langle 0| + b\,|0\rangle\langle 1| + c\,|1\rangle\langle 0| + d\,|1\rangle\langle 1|$$

In particular, $|0\rangle\langle 0|, |0\rangle\langle 1|, |1\rangle\langle 0|, |1\rangle\langle 1|$ forms a basis for the vector space $\mathcal{V} \otimes \mathcal{V}^\star$ of linear maps on $\mathcal{V}$. Note also that $\langle w|v\rangle = \operatorname{Tr}|v\rangle\langle w|$.

1.8. Projection operators

Suppose that $|v\rangle$ is a normalised vector, so $\langle v|v\rangle = 1$. Then, $\Pi_v = |v\rangle\langle v|$ is the *projection operator* onto v, satisfying $\Pi_v\Pi_v = \Pi_v$ and $\Pi_v^\dagger = \Pi_v$. In Dirac notation, one can see that

$$\Pi_v\Pi_v = |v\rangle\langle v|\,|v\rangle\langle v| = |v\rangle\,\langle v|v\rangle\,\langle v| = |v\rangle\langle v| = \Pi_v$$

If $|a\rangle$ is orthogonal to $|v\rangle$, then $\Pi_v |a\rangle = |v\rangle\langle v|a\rangle = 0$. Therefore, $\Pi_v |x\rangle$ is the vector obtained by projection of $|x\rangle$ onto the one-dimensional subspace of $\mathcal{V}$ spanned by $|v\rangle$.

Now suppose $\mathcal{E}$ is any linear subspace of some vector space $\mathcal{V}$, and $|e_1\rangle, \dots, |e_d\rangle$ is any orthonormal basis of $\mathcal{E}$. Then,

$$\Pi_{\mathcal{E}} = |e_1\rangle\langle e_1| + \cdots + |e_d\rangle\langle e_d|$$

is the projection operator into $\mathcal{E}$. This property can be checked by extending $|e_1\rangle, \dots, |e_d\rangle$ into an orthonormal basis of $\mathcal{V}$.

Note that if $|x\rangle = A|v\rangle$, then $\langle x| = (A|v\rangle)^\dagger = |v\rangle^\dagger A^\dagger = \langle v| A^\dagger$. Therefore, when constructing inner products, we can write $\langle a|M|b\rangle$ as $\langle a|x\rangle$ or $\langle y|b\rangle$ where $|x\rangle = M|b\rangle$ or $|y\rangle = M^\dagger |a\rangle$ (so that we have $\langle y| = \langle a| M$).

1.9. Tensor products of linear maps

Suppose A, B are linear maps $\mathbb{C}^2 \to \mathbb{C}^2$. Then, we define $A \otimes B : \mathbb{C}^2 \otimes \mathbb{C}^2 \to \mathbb{C}^2 \otimes \mathbb{C}^2$ by its action on the basis $(A \otimes B)|i\rangle|j\rangle = A|i\rangle B|j\rangle$. In particular, for product vectors we obtain $(A \otimes B)(|v\rangle|w\rangle) = A|v\rangle \otimes B|w\rangle$.

The 4×4 matrix of components of $A \otimes B$ has a simple block form, which can be seen by writing down its action on basis states.

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}; \quad B = \begin{pmatrix} p & q \\ r & s \end{pmatrix} \implies A \otimes B = \begin{pmatrix} aB & bB \\ cB & dB \end{pmatrix} = \begin{pmatrix} ap & aq & bp & bq \\ ar & as & br & bs \\ cp & cq & dp & dq \\ cr & cs & dr & ds \end{pmatrix}$$

Note that $A \otimes I$ and $I \otimes A$ can be thought of as acting only on one of the subspaces. Consider $|\Phi\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$, and define A as above. Then,

$$\begin{aligned}
(A \otimes I)|\Phi\rangle &= \frac{1}{\sqrt{2}}[(A|0\rangle)|0\rangle + (A|1\rangle)|1\rangle] \\
&= \frac{1}{\sqrt{2}}[(a|0\rangle + c|1\rangle)|0\rangle + (b|0\rangle + d|1\rangle)|1\rangle] \\
&= \frac{1}{\sqrt{2}}[a|00\rangle + b|01\rangle + c|10\rangle + d|11\rangle] \\
(I \otimes A)|\Phi\rangle &= \frac{1}{\sqrt{2}}[|0\rangle(A|0\rangle) + |1\rangle(A|1\rangle)] \\
&= \frac{1}{\sqrt{2}}[|0\rangle(a|0\rangle + c|1\rangle) + |1\rangle(b|0\rangle + d|1\rangle)] \\
&= \frac{1}{\sqrt{2}}[a|00\rangle + c|01\rangle + b|10\rangle + d|11\rangle]
\end{aligned}$$

1.10. Third postulate: physical evolution of quantum systems

The third postulate of quantum mechanics is that any physical finite-time evolution of a closed quantum system is represented by a unitary operation on the corresponding vector space of states. Recall that the following are equivalent for a linear operator U:

- U is unitary, so $U^{-1} = U^\dagger$;
- U maps an orthonormal basis to an orthonormal set of vectors;
- the columns (or rows) of U form an orthonormal set of vectors.

If a system is in a state $|\psi(t_1)\rangle$ at a time t_1 and later in a state $|\psi(t_2)\rangle$ at a time t_2, then $|\psi(t_2)\rangle = U(t_1, t_2)\,|\psi(t_1)\rangle$ for some unitary map $U(t_1, t_2)$ which depends only on t_1, t_2. This operator is derived from the *Schrödinger equation*, which is

$$i\hbar\frac{\partial}{\partial t}\,|\psi(t)\rangle = H\,|\psi(t)\rangle$$

where H is a self-adjoint operator known as the *Hamiltonian*. In particular, if H is time-independent, we have

$$U(t_1, t_2) = e^{-\frac{i}{\hbar}H(t_2-t_1)}$$

In the more general case,

$$U(t_1, t_2) = e^{-\frac{i}{\hbar}\int_{t_1}^{t_2} H(t)\mathrm{d}t}$$

The unitary evolution of a closed system is deterministic.

1.11. Partial inner products

A vector $|v\rangle \in \mathcal{V}$ defines a linear map $\mathcal{V} \otimes \mathcal{W} \to \mathcal{W}$ called the *partial inner product* with $|v\rangle$, defined on the basis $|e_i\rangle\,|f_j\rangle$ of $\mathcal{V} \otimes \mathcal{W}$ by $|e_i\rangle\,|f_j\rangle \mapsto \langle v|e_i\rangle\,|f_j\rangle$. Similarly, for any $|w\rangle \in \mathcal{W}$, we obtain a partial inner product $\mathcal{V} \otimes \mathcal{W} \to \mathcal{V}$. If $\mathcal{V}, \mathcal{W}$ are isomorphic, we must specify which partial inner product is intended.

1.12. Fourth postulate: quantum measurement

Consider a system S with state space $\mathcal{V}$, and let A be an observable. A can be written as its *spectral projection* $A = \sum_k a_k P_k$ where $A\,|\varphi_k\rangle = a_k\,|\varphi_k\rangle$. If a_k is nondegenerate, $P_k = |\varphi_k\rangle\langle\varphi_k|$. If a_k is degenerate of multiplicity m, then $P_k = \sum_{i=1}^m |\varphi_k^i\rangle\langle\varphi_k^i|$.

The fourth postulate is that when an observable is measured, the resulting measurement will be an eigenvalue a_j, with probability $p(a_j) = \langle\psi|P_j|\psi\rangle$. Then, $|\psi\rangle$ is replaced with the post-measurement state

$$\frac{P_j\,|\psi\rangle}{\sqrt{p(a_j)}}$$

This is known as *Born's rule*. Such a measurement is called a *projective measurement* (or sometimes a *von Neumann measurement*), since the post-measurement state is given by a projection operator.

Suppose A, B are operators that do not commute, so $[A, B] = AB - BA \neq 0$. Then, the measurement of A will influence the outcome probabilities of a subsequent measurement of B. For instance, suppose $|\psi\rangle = |+\rangle, A = \sigma_z, B = \sigma_x$.

1.13. Complete and incomplete projective measurements

Let $|\psi\rangle \in \mathcal{V}$ be a state in a state space of dimension n. Let $\mathcal{B} = \{|e_i\rangle\}$ be a set of n orthogonal basis vectors for $\mathcal{V}$. Then $|\psi\rangle = \sum a_j |e_j\rangle$ where $a_k = \langle e_k|\psi\rangle$. If the outcomes of a measurement are the indices of the basis vectors $j = 1, \dots, n$, we have $p(j) = \langle\psi|P_j|\psi\rangle$ where $P_j = |e_j\rangle\langle e_j|$. Therefore, $p(j) = |\langle\psi|e_j\rangle|^2 = |a_j|^2$. If the outcome is j, the post-measurement state is

$$\frac{P_j|\psi\rangle}{\sqrt{p(j)}} = \frac{|e_j\rangle\langle e_j|\psi\rangle}{\sqrt{p(j)}} = |e_j\rangle$$

Hence the state collapses to a basis vector. Taking another measurement immediately in the same basis, we obtain the result j with probability 1. Such a measurement is called a *complete* projective measurement; it is called complete as all P_j are of rank 1. When we measure a state $|\psi\rangle$ in a basis, it is often helpful to consider an orthogonal decomposition of $\mathcal{V}$ using the basis vectors.

Conversely, an *incomplete* projective measurement corresponds to an arbitrary orthogonal decomposition of $\mathcal{V}$. Consider a decomposition of $\mathcal{V}$ into d mutually orthogonal subspaces $\mathcal{E}_1, \dots, \mathcal{E}_d$, so $\mathcal{V} = \mathcal{E}_1 \oplus \dots \oplus \mathcal{E}_d$, and $\dim \mathcal{V} = \sum \dim \mathcal{E}_j$. Let Π_i be a projection operator onto $\mathcal{E}_i$. Since the spaces are mutually orthogonal, $\Pi_i\Pi_j = \delta_{ij}\Pi_i$. Consider a measurement with outcomes $1, \dots, d$ representing a particular subspace. The probability of observing outcome i is $\langle\psi|\Pi_i|\psi\rangle$. If the outcome is i, $|\psi\rangle$ is replaced with $\frac{\Pi_i|\psi\rangle}{\sqrt{p(i)}}$. In this case, the Π_i are no longer rank 1 projection operators. If $\mathcal{E}_i$ has basis $\{|f_j\rangle\}$, we can write $\Pi_i = \sum |f_j\rangle\langle f_j|$.

Incomplete projective measurement is a generalisation of complete projective measurement. One can refine an incomplete measurement into a complete measurement by first considering a complete measurement, and then summing the relevant outcome probabiilities to obtain a description of the incomplete measurement probabilities. Let $\left\{\left|e_k^{(j)}\right\rangle\right\}_{k=1}^{d_j}$ be a basis for $\mathcal{E}_j$ for each j. Then $\mathcal{V} = \bigoplus_{i=1}^d \mathcal{E}_j$ has orthonormal basis $\left\{\left|e_k^{(j)}\right\rangle\right\}_{j,k}$. Then, $\left\langle e_i^{(k_1)}\middle|e_j^{(k_2)}\right\rangle = \delta_{ij}\delta_{k_1k_2}$.

Consider a two-bit string b_1b_2. The *parity* of this string is $b_1 \oplus b_2$, where $\oplus$ represents addition modulo 2. Consider the orthogonal decomposition of $\mathcal{V}$ into $\mathcal{E}_0 \oplus \mathcal{E}_1$, where $\mathcal{E}_0 = \operatorname{span}\{|00\rangle, |11\rangle\}$ is the even parity subspace, and $\mathcal{E}_1 = \operatorname{span}\{|01\rangle, |10\rangle\}$ is the odd parity subspace. The outcomes of an incomplete measurement are then the labels 0 and 1 of the subspaces $\mathcal{E}_0$ and $\mathcal{E}_1$. Note that $\{|00\rangle, |01\rangle, |10\rangle, |11\rangle\}$ is a complete orthonormal basis for

$\mathcal{V}$, so we can consider the complete projective measurement. $\langle\psi|P_{00}|\psi\rangle$ is the probability of outcome 00 for the complete measurement, where $P_{00} = |00\rangle\langle 00|$. For the incomplete measurement, $p(0) = \langle\psi|\Pi_0|\psi\rangle$ is the probability of outcome 0, where $\Pi_0 = P_{00} + P_{11}$. So $p(0) = \langle\psi|P_{00}|\psi\rangle + \langle\psi|P_{11}|\psi\rangle$.

1.14. Extended Born rule

Let S_1, S_2 be quantum systems with state spaces $\mathcal{V}, \mathcal{W}$ with dimensions m, n, and we consider the composite system S_1S_2. Let $\{|e_i\rangle\}$ be a complete orthonormal basis of $\mathcal{V}$, and let $\{|f_j\rangle\}$ be a complete orthonormal basis of $\mathcal{W}$. Suppose the composite system is in an initial state $|\psi\rangle = \sum a_{ij}\,|e_i\rangle\,|f_j\rangle$. Suppose now that we want to measure $|\psi\rangle$ in the basis $\{|e_i\rangle\}$; this amounts to an incomplete measurement with subspaces $\mathcal{E}_i = \operatorname{span}\{|e_i\rangle \otimes |\varphi\rangle \mid |\varphi\rangle \in \mathcal{W}\}$ for $1 \le i \le m$. The outcomes of such a measurement are $\{1, \dots, m\}$, and the $\mathcal{E}_i$ are mutually orthogonal. The probability of a given outcome is $p(k) = \langle\psi|P_k \otimes I|\psi\rangle$, where $P_k = |e_k\rangle\langle e_k|$. Hence,

$$p(k) = \left(\sum a^\star_{i'j'}\,\langle e'_i|\,\langle f'_j|\right)(|e_k\rangle\langle e_k| \otimes I)\left(\sum a_{ij}\,|e_i\rangle\,|f_j\rangle\right) = \sum_{j=1}^{n} a^\star_{kj}a_{kj}$$

If the outcome is k, then the post-measurement state is given by

$$|\psi_{\text{after}}\rangle = \frac{(P_k \otimes I)\,|\psi\rangle}{p(k)} = \frac{\sum_j a_{kj}\,|e_k\rangle\,|f_j\rangle}{\sqrt{\sum_j |a_{kj}|^2}}$$

Using partial inner products, one can show that $|\psi_{\text{after}}\rangle$ is normalised. These rules are referred to as the *extended Born rule.*

Consider a quantum system S with state space $\mathcal{V}$. A measurement relative to any basis $\mathcal{C}$ can be performed by first performing a unitary operator, then performing a measurement in a fixed basis $\mathcal{B}$. Let $\mathcal{B} = \{|e_i\rangle\}$, and $\mathcal{C} = \{|e'_i\rangle\}$. Let U be a unitary operator such that $|e'_i\rangle = U\,|e_i\rangle$. Then, $U^\dagger = U^{-1}$ has the property that $U^{-1}\,|e'_i\rangle = |e_i\rangle$. Suppose we have a state $|\psi\rangle \in \mathcal{V}$. Let $|\psi\rangle = \sum c_i\,|e'_i\rangle$. Applying U^{-1} to $|\psi\rangle$, we obtain $U^{-1}\,|\psi\rangle = \sum c_i\,|e_i\rangle$ by linearity. We can then measure $|\psi'\rangle = U^{-1}\,|\psi\rangle$ in the basis $\mathcal{B}$. By the Born rule, $p(i) = \langle\psi'|P_i|\psi'\rangle = \langle\psi|\,UP_iU^\dagger\,|\psi\rangle$ where $P_i = |e_i\rangle\langle e_i|$, as we are performing a complete projective measurement. If the outcome is i, then the post-measurement state is $|\psi'_{\text{after}}\rangle = \frac{P_i|\psi'\rangle}{p(i)}$.

1.15. Standard measurement on multi-qubit systems

Consider a system of n qubits. The state space is $(\mathbb{C}^2)^{\otimes n}$. The *computational basis* or *standard basis* is $\mathcal{B} = \{|i_1 \dots i_n\rangle \mid i_j \in \{0,1\}\}$. The labels of the elements of the standard basis are labelled by bit strings of length n.

Suppose we are measuring a subset of k qubits of the n-qubit system. Let $n = 3$, and let

$$|\psi\rangle = \frac{i}{2}\,|000\rangle + \frac{1+i}{2\sqrt{2}}\,|001\rangle - \frac{1}{2}\,|101\rangle + \frac{3}{10}\,|110\rangle - \frac{2i}{5}\,|111\rangle$$

The standard measurement of any of the three qubits will always have the outcome zero or one. Suppose we perform a standard measurement on the first qubit. By the extended Born rule, we obtain

$$p^{(1)}(1) = \langle\psi| P_1 \otimes I \otimes I |\psi\rangle = \langle\psi| (|1\rangle\langle 1| \otimes I \otimes I) |\psi\rangle = \frac{1}{4} + \frac{9}{100} + \frac{4}{25} = \frac{1}{2}$$

If we measure the outcome 1, the post-measurement state is $|\psi_{\text{after}}\rangle = \frac{(P_1 \otimes I \otimes I)|\psi\rangle}{\sqrt{p^{(1)}(1)}}$.

1.16. Reliably distinguishing states

Note that the measurement postulate implies that states with guaranteed (with probability 1) different measurement outcomes always lie in mutually orthogonal subspaces. We say that two states are *reliably distinguishable* if there exists a measurement which outputs two distinct outcomes with probability 1 when applied to the two states. Therefore, two states $|\psi\rangle, |\varphi\rangle$ are reliably distinguishable if and only if they are orthogonal, so $\langle\psi|\varphi\rangle = 0$.

Let $|\psi\rangle$ and $|\varphi\rangle$ be orthogonal. Let $\mathcal{B} = \{|\psi\rangle, |f_1\rangle, \dots, |f_{m-1}\rangle\}$ be a complete orthonormal basis for $\mathcal{V}$. Then $\langle\psi|f_j\rangle = 0$ and $\langle f_j|f_k\rangle = \delta_{jk}$. Measuring $|\psi\rangle$ in this basis, $p(1) = \langle\psi| P_1 |\psi\rangle$ where $P_1 = |\psi\rangle\langle\psi|$, so the probability is 1. Measuring $|\varphi\rangle$ in this basis, $p(1) = \langle\psi|\varphi\rangle\langle\varphi|\psi\rangle = 0$. This is an example of a measurement which can reliably distinguish $|\psi\rangle$ and $|\varphi\rangle$.

Vectors $|v\rangle = |\psi\rangle$ and $|v'\rangle = e^{i\theta}|\psi\rangle$ are not distinguishable. For any measurement, the probability of obtaining a particular outcome when measuring $|v\rangle$ is always the same as the probability when measuring $|v'\rangle$.

2. Quantum states as information carriers

2.1. Using higher Hilbert spaces

Quantum information is encoded in the states of a quantum system. Classical information is encoded in states chosen from an orthonormal set, since all distinct classical messages can be distinguished. Given a quantum system S and a quantum state $|\psi\rangle$, we can perform this sequence of operations.

- (ancilla) Consider an auxiliary system A in a fixed state $|A\rangle \in \mathcal{V}_A$. The composite system SA has vector space $\mathcal{V}_S \otimes \mathcal{V}_A$. The initial joint state is $|\psi\rangle\,|A\rangle$. This results in an embedding of quantum information in a higher dimensional space.
- (unitary) Consider the action of a unitary operator U on SA (or on S), modelling the time evolution of the quantum system.
- (measure) We can perform measurements on SA (or on S). The post-measurement state of S is retained, and the auxiliary system A is discarded.

This process is sometimes known as 'going to the church of the higher Hilbert space'. The presence of the ancilla allows for entanglement with other quantum systems.

2.2. No-cloning theorem

Classically, information can be easily copied by measuring all relevant information and reproducing it. Quantum copying involves three systems:

- a system A containing some quantum information to be copied;
- a system B with $\mathcal{V}_B \simeq \mathcal{V}_A$ initially in some fixed state $|0\rangle$ where the information is to be copied;
- a system M which represents any physical machinery in some 'ready' state $|M_0\rangle$ required for performing the copy.

The initial state of this composite system ABM is $|\psi\rangle\,|0\rangle\,|M_0\rangle$. Note that the $|\psi\rangle$ and $|0\rangle\,|M_0\rangle$ are *uncorrelated* in this state, as we are using the tensor product to combine them. Suppose that the cloning process is performed using some unitary operator U, so $U\,|\psi_A\rangle\,|0\rangle\,|M_0\rangle = |\psi_A\rangle\,|\psi_B\rangle\,|M_\psi\rangle$. This cloning process may be required to work either for all states of A, or for some subset of A.

Theorem. Let $\mathcal{S}$ be any set of states of the system A that contains at least one pair of distinct non-orthogonal states. Then there does not exist any unitary operator U that clones all states in $\mathcal{S}$.

Proof. Let $|\xi\rangle\,, |\eta\rangle$ be distinct non-orthogonal states in $\mathcal{S}$, so $\langle\xi|\eta\rangle \neq 0$. Suppose such a unitary operator U exists. Then, we must have

$$U\,|\xi_A\rangle\,|0_B\rangle\,|M_0\rangle = |\xi_A\rangle\,|\xi_B\rangle\,|M_\xi\rangle\,; \quad U\,|\eta_A\rangle\,|0_B\rangle\,|M_0\rangle = |\eta_A\rangle\,|\eta_B\rangle\,|M_\eta\rangle$$

Unitary operators preserve inner products. Hence,

$$\langle \xi_A | \eta_A \rangle \langle 0_B | 0_B \rangle \langle M_0 | M_0 \rangle = \langle \xi_A | \eta_A \rangle \langle \xi_B | \eta_B \rangle \langle M_\xi | M_\eta \rangle$$

Hence, $\langle \xi | \eta \rangle = (\langle \xi | \eta \rangle)^2 \langle M_\xi | M_\eta \rangle$. By taking the absolute value, $|\langle \xi | \eta \rangle| = |\langle \xi | \eta \rangle|^2 |\langle M_\xi | M_\eta \rangle|$. Since $\xi \neq \eta$, we must have $0 < |\langle \xi | \eta \rangle| < 1$, and $0 \leq |\langle M_\xi | M_\eta \rangle| \leq 1$. Therefore, $1 = |\langle \xi | \eta \rangle| |\langle M_\xi | M_\eta \rangle| < 1$, which is a contradiction. □

If quantum cloning were possible, superluminal (indeed, instantaneous) communication would also be possible. Suppose we have a state $|\psi_{AB}^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle) \in \mathbb{C}^2 \otimes \mathbb{C}^2$. Let A, B be the entangled parts of this quantum state, and suppose that we send qubit A to Alice and B to Bob, far apart from each other.

If we want to send the bit 'yes' from Alice to Bob, we measure the qubit A in the basis $\{|0\rangle, |1\rangle\}$, which gives outcomes $0, 1$ with probability $\frac{1}{2}$. If the outcome is 0, the final state of B is $|0\rangle$, and if the outcome is 1, the final state of B is $|1\rangle$. If we want to send 'no', we instead measure A in the basis $\{|+\rangle, |-\rangle\}$, which gives the outcomes $+, -$ with probability $\frac{1}{2}$. Similarly, the final state of B is $|+\rangle$ or $|-\rangle$.

We claim that these 'yes' ($|0\rangle, |1\rangle$) and 'no' ($|+\rangle, |-\rangle$) *preparations* of qubit B are indistinguishable by Bob with any local action on the qubit. That is, they each give exactly the same probability distribution of outcomes of any measurement. In fact, the distribution matches the prior distribution before qubit A was measured.

Let Π_i be the projection operator for outcome i on qubit B. Suppose that 'yes' was sent. Then,

$$p_{\text{yes}}(i) = \frac{1}{2} \langle 0 | \Pi_i | 0 \rangle + \frac{1}{2} \langle 1 | \Pi_i | 1 \rangle = \frac{1}{2} \operatorname{Tr} [\Pi_i (|0\rangle\langle 0| + |1\rangle\langle 1|)] = \frac{1}{2} \operatorname{Tr} \Pi_i$$

In the 'no' case,

$$p_{\text{no}}(i) = \frac{1}{2} \langle + | \Pi_i | + \rangle + \frac{1}{2} \langle - | \Pi_i | - \rangle = \frac{1}{2} \operatorname{Tr} [\Pi_i (|+\rangle\langle +| + |-\rangle\langle -|)] = \frac{1}{2} \operatorname{Tr} \Pi_i$$

These probability distributions match.

Suppose that cloning were possible. We clone the qubit B multiple times after the message was sent, to produce one of the states $|0\rangle \dots |0\rangle, |1\rangle \dots |1\rangle, |+\rangle \dots |+\rangle, |-\rangle \dots |-\rangle$. We now measure each qubit in the basis $|0\rangle, |1\rangle$ separately. If the 'yes' message was sent, all measurements will result in 0 or 1. If 'no' was sent, it is possible that two measurements would differ. In expectation, half of the measurements would result in the outcome 0 and half would result in the outcome 1. Therefore, the 'yes' and 'no' errors can be distinguished with probability of error 2^{-N+1} if we make N copies of B.

2.3. Distinguishing non-orthogonal states

Suppose you know a state $|\psi\rangle$ has state $|\alpha_0\rangle$ or $|\alpha_1\rangle$ with probability $\frac{1}{2}$, where $\langle \alpha_0 | \alpha_1 \rangle \neq 0$. Since the states are non-orthogonal, we cannot perfectly distinguish the states, but must allow some error rate. The simplest possibility is to not make a measurement and guess randomly; in which case, the guess is correct with probability $\frac{1}{2}$.

Suppose we append an auxiliary system $|A\rangle$ to $|\alpha_i\rangle$. Note that $\langle A| \langle \alpha_i|\alpha_i\rangle |A\rangle = \langle \alpha_i|\alpha_i\rangle$ as $|A\rangle$ is normalised. If we apply a unitary operator U to $|\alpha_i\rangle$ then perform a projective measurement in the basis $\{\Pi_0, \Pi_1\}$, our action corresponds to simply performing a measurement $\Pi_0' = U^\dagger \Pi_0 U$ or $\Pi_1' = U^\dagger \Pi_1 U$, which leads to the same probabilities of outcomes. Indeed,

$$p(i) = \langle U\xi|\Pi_i|U\xi\rangle = \langle \xi|U^\dagger \Pi_i U|\xi\rangle = \langle \xi|\Pi_i'|\xi\rangle$$

Therefore, in this particular problem, we gain no benefit from moving to a larger Hilbert space or applying unitary operators.

We now describe the *state estimation* or *state discrimination* process. We will consider a two-outcome measurement $\{\Pi_0, \Pi_1\}$, where $\Pi_0 + \Pi_1 = I$. The average success probability is

$$\begin{aligned} p_S(\Pi_0, \Pi_1) &= \frac{1}{2}\mathbb{P}\left(0 \mid |\psi\rangle = |\alpha_0\rangle\right) + \frac{1}{2}\mathbb{P}\left(1 \mid |\psi\rangle = |\alpha_1\rangle\right) \\ &= \frac{1}{2}\langle \alpha_0|\Pi_0|\alpha_0\rangle + \frac{1}{2}\langle \alpha_1|\Pi_1|\alpha_1\rangle \\ &= \frac{1}{2} + \frac{1}{2}\operatorname{Tr}\left[\Pi_0(|\alpha_0\rangle\langle\alpha_0| - |\alpha_1\rangle\langle\alpha_1|)\right] \end{aligned}$$

as $\operatorname{Tr}(A\,|\psi\rangle\langle\psi|) = \langle \alpha|A|\alpha\rangle$. The optimal choice of measurement maximises the average success probability p_S. Note that $\Delta = |\alpha_0\rangle\langle\alpha_0| - |\alpha_1\rangle\langle\alpha_1|$ is self-adjoint, and we can write $p_S = \frac{1}{2} + \frac{1}{2}\operatorname{Tr}(\Pi_0\Delta)$. Therefore, the eigenvalues of Δ are real, and the eigenvectors form an orthonormal basis. For a state $|\beta\rangle$ orthogonal to both $|\alpha_0\rangle$ and $|\alpha_1\rangle$, we have $\Delta\,|\beta\rangle = 0$. Therefore, Δ acts nontrivially only in the vector space spanned by $|\alpha_0\rangle$ and $|\alpha_1\rangle$, and hence has at most two nonzero eigenvalues, and its eigenvectors lie in $\operatorname{span}\{|\alpha_0\rangle, |\alpha_1\rangle\}$.

Now, $\operatorname{Tr}\Delta = 0$ so the eigenvalues are δ and $-\delta$ for some $\delta \in \mathbb{R}$. Let $|p\rangle$ be the eigenvector for δ, and $|m\rangle$ be the eigenvector for $-\delta$, so $\langle p|m\rangle = 0$. We can write Δ in its spectral decomposition, giving $\Delta = \delta\,|p\rangle\langle p| - \delta\,|m\rangle\langle m|$.

Let $|\alpha_0^\perp\rangle \in \operatorname{span}\{|\alpha_0\rangle, |\alpha_1\rangle\}$ be a normalised vector such that $\langle \alpha_0^\perp|\alpha_0\rangle = 0$. Then, $\{|\alpha_0\rangle, |\alpha_0^\perp\rangle\}$ is an orthonormal basis. Hence, we can write $|\alpha_1\rangle = c_0\,|\alpha_0\rangle + c_1\,|\alpha_0^\perp\rangle$. In this basis,

$$\Delta = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} -|c_0|^2 & -c_0c_1^\star \\ -c_0^\star c_1 & -|c_1|^2 \end{pmatrix} = \begin{pmatrix} 1-|c_0|^2 & -c_0c_1^\star \\ -c_0^\star c_1 & -|c_1|^2 \end{pmatrix} = \begin{pmatrix} |c_1|^2 & -c_0c_1^\star \\ -c_0^\star c_1 & -|c_1|^2 \end{pmatrix}$$

which has eigenvalues $\delta = |c_1|, -\delta = -|c_1|$. Since $|c_0| = |\langle \alpha_0|\alpha_1\rangle| = \cos\theta$ where $\theta \geq 0$, we have $\delta = \sin\theta$. Then,

$$\begin{aligned} p_S(\Pi_0, \Pi_1) &= \frac{1}{2} + \frac{1}{2}\operatorname{Tr}(\Pi_0\Delta) \\ &= \frac{1}{2} + \frac{1}{2}\operatorname{Tr}(\Pi_0[\sin\theta\,|p\rangle\langle p| - \sin\theta\,|m\rangle\langle m|]) \\ &= \frac{1}{2} + \frac{\sin\theta}{2}[\langle p|\Pi_0|p\rangle - \langle m|\Pi_0|m\rangle] \end{aligned}$$

Note that for any $|\varphi\rangle$, we have $0 \leq \langle\varphi|\Pi|\varphi\rangle \leq 1$, so the measurement is maximised when $\langle p|\Pi_0|p\rangle = 1$ and $\langle m|\Pi_0|m\rangle = 0$. We therefore define $\Pi_0 = |p\rangle\langle p|$. Then, the optimal average success probability is

$$p_S^\star = \frac{1}{2} + \frac{\sin\theta}{2}$$

Theorem (Holevo–Helstrom theorem for pure states). Let $|\alpha_0\rangle, |\alpha_1\rangle$ be equally likely states, with $|\langle\alpha_0|\alpha_1\rangle| = \cos\theta$, $\theta \geq 0$. Then, the probability p_S of correctly identifying the state by any quantum measurement satisfies

$$p_S \leq \frac{1}{2} + \frac{\sin\theta}{2}$$

and this bound can be attained.

In the case of orthogonal states, the theorem implies that $p_S \leq 1$ and the bound can be attained, which was shown before.

2.4. No-signalling principle

Suppose we have a possibly entangled state $|\phi_{AB}\rangle \in \mathcal{V}_A \otimes \mathcal{V}_B$ shared between two agents Alice (A) and Bob (B). Suppose we perform a complete projective measurement on $|\phi_A\rangle$. By the extended Born rule, each measurement outcome will lead to an instantaneous change of $|\phi_B\rangle$. If this change in state could be detected by measuring $|\phi_B\rangle$, instantaneous communication between A and B would be possible.

Consider $|\phi_{AB}^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$. Suppose qubit A is measured in the standard basis $\{|0\rangle, |1\rangle\}$.

outcome	probability	post-measurement state	final state of B
0	$\frac{1}{2}$	$\lvert 00\rangle$	$\lvert 0\rangle$
1	$\frac{1}{2}$	$\lvert 11\rangle$	$\lvert 1\rangle$

Suppose qubit B is subsequently measured in $\{|b_0\rangle, |b_1\rangle\}$. If B is in the state $|0\rangle$, we can write $|0\rangle = c_0|b_0\rangle + c_1|b_1\rangle$, and $p_{|0\rangle}(i) = |c_i|^2 = |\langle b_i|0\rangle|^2$. If B is in the state $|1\rangle$, we write $|1\rangle = d_0|b_0\rangle + d_1|b_1\rangle$, and $p_{|1\rangle}(i) = |d_i|^2 = |\langle b_i|1\rangle|^2$. Therefore, $p(i) = \frac{1}{2}|\langle b_i|0\rangle|^2 + \frac{1}{2}|\langle b_i|1\rangle|^2 = \frac{1}{2}$. The two outcomes for this measurement are equally likely, regardless of the choice of complete orthonormal basis $\{|b_0\rangle, |b_1\rangle\}$.

Suppose instead A is not measured, but we perform the same measurement on B. The initial state is $|\phi_{AB}^+\rangle$, so by the extended Born rule, $p(i) = \langle\phi_{AB}^+|(I_A \otimes |b_i\rangle\langle b_i|)|\phi_{AB}^+\rangle = \frac{1}{2}$. We can therefore not detect through measuring B whether a measurement was performed at A. This is the no-signalling principle.

We now prove the more general case. Let $|\phi_{AB}\rangle \in \mathcal{V}_A \otimes \mathcal{V}_B$ be an arbitrary possibly entangled state.

Suppose we measure B in a complete orthonormal basis $\{|b\rangle\}_{b=1}^{\dim \mathcal{V}_B}$, which is a complete projective measurement on B. Let $\{|a\rangle\}_{a=1}^{\dim \mathcal{V}_A}$ be a complete orthonormal basis for $\mathcal{V}_A$. Then, expanding $|\phi_{AB}\rangle$, in this basis, we can write $|\phi_{AB}\rangle = \sum_{a,b} c_{ab} |a\rangle |b\rangle$. We obtain outcome b with probability $p(b) = \langle\phi_{AB}|(I_A \otimes P_b)|\phi_{AB}\rangle = \sum_{a=1}^{\dim \mathcal{V}_A} |c_{ab}|^2$. The post-measurement state is $|\phi'_{AB}\rangle$.

Suppose that we first measure A in a complete orthonormal basis $\{|a\rangle\}_{a=1}^{\dim \mathcal{V}_A}$, and then perform the measurement $\{|b\rangle\}_{b=1}^{\dim \mathcal{V}_B}$ on B. The outcome of the first measurement is a with probability $p(a) = \langle\phi_{AB}|(P_a \otimes I_B)|\phi_{AB}\rangle = \sum_{b=1}^{\dim \mathcal{V}_B} |c_{ab}|^2$. We denote the post-measurement state of the joint system by $|\phi''_{AB}\rangle = \frac{(P_a\otimes I_B)|\phi_{AB}\rangle}{\sqrt{p(a)}}$. Then, the outcome of the second measurement is b with probability

$$\begin{aligned} p(a \mid b) &= \langle\phi''_{AB}|(I_A \otimes P_b)|\phi''_{AB}\rangle \\ &= \frac{1}{p(a)} \langle\phi_{AB}|(P_a \otimes I_B)(I_A \otimes P_b)(P_a \otimes I_B)|\phi_{AB}\rangle \\ &= \frac{1}{p(a)} \langle\phi_{AB}|(P_a \otimes P_b)|\phi_{AB}\rangle \\ p(a,b) = p(a)p(a \mid b) &= \langle\phi_{AB}|(P_a \otimes P_b)|\phi_{AB}\rangle = |c_{ab}|^2 \end{aligned}$$

Hence $p(b) = \sum_{a=1}^{\dim \mathcal{V}_A} |c_{ab}|^2$, which is exactly the distribution we obtained when no measurement on A was performed. This proves the no-signalling principle.

2.5. The Bell basis

Let $\mathbb{C}^2 \otimes \mathbb{C}^2$ model a quantum system representing the spins of two electrons. Consider $|\phi^+_{AB}\rangle = \frac{1}{2}(|00\rangle + |11\rangle) \in \mathbb{C}^2 \otimes \mathbb{C}^2$. This is a *maximally entangled state*; we have information about the whole system, but no information about the individual states.

$$|\phi^\pm_{AB}\rangle = \frac{1}{\sqrt{2}}(|00\rangle \pm |11\rangle); \quad |\psi^\pm_{AB}\rangle = \frac{1}{\sqrt{2}}(|01\rangle \pm |10\rangle)$$

$\{|\phi^\pm_{AB}\rangle, |\psi^\pm_{AB}\rangle\}$ forms a complete orthonormal basis of $\mathbb{C}^2 \otimes \mathbb{C}^2$. This is called the *Bell basis*. The basis vectors are sometimes known as *EPR states*, after Einstein, Podolsky, and Rosen.

One bit of classical information can be encoded in a single qubit, and two bits can be encoded in a pair of qubits in the Bell basis. The Bell states have a *parity* 0 or 1, representing parallel $\{|\phi^\pm\rangle\}$ or antiparallel $\{|\psi^\pm\rangle\}$ spins. The states also have a *phase*, which can be positive $\{|\phi^+\rangle, |\psi^+\rangle\}$ or negative $\{|\phi^-\rangle, |\psi^-\rangle\}$. For example, we can encode the classical message 01 using the state $|\phi^-\rangle$.

We can perform a complete projective measurement on both qubits in the Bell basis to recover the encoded information with certainty. For instance, $P_{00} = |\phi^+\rangle\langle\phi^+|$. If we prepare a pair of electrons $|\phi\rangle$ in the state $|\phi^-\rangle$ for example, we obtain $p(00) = p(10) = p(11) = 0$ and $p(01) = 1$.

2.6. Superdense coding

Suppose Alice wants to send a classical message to Bob. Two bits of classical information can be sent reliably via a single qubit, provided that Alice and Bob share an entangled state, using *superdense coding* or *quantum dense coding*. Let

$$X = \sigma_x; \quad Z = \sigma_z; \quad Y = i\sigma_y = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

One can check that the Bell basis vectors satisfy

$$\begin{aligned} |\phi^+\rangle &= (I \otimes I)\,|\phi^+\rangle = (I \otimes I)\,|\phi^+\rangle \\ |\phi^-\rangle &= (Z \otimes I)\,|\phi^+\rangle = (I \otimes Z)\,|\phi^+\rangle \\ |\psi^+\rangle &= (X \otimes I)\,|\phi^+\rangle = (I \otimes X)\,|\phi^+\rangle \\ |\psi^-\rangle &= (Y \otimes I)\,|\phi^+\rangle = -(I \otimes Y)\,|\phi^+\rangle \end{aligned}$$

Suppose we have shared the entangled Bell state $|\phi^+_{AB}\rangle$ between Alice and Bob. The superdense coding protocol is

Alice's message	local action on A	final state of AB
00	I	$\lvert\phi^+\rangle$
01	Z	$\lvert\phi^-\rangle$
10	X	$\lvert\psi^+\rangle$
11	Y	$\lvert\psi^-\rangle$

Then, Alice sends qubit A to Bob, so Bob has the entire state AB. Bob performs a Bell measurement, which distinguishes between the four Bell states, thus recovering Alice's message. Since the state is maximally entangled, an eavesdropper who may intercept Alice's transmission cannot recover any part of the message.

2.7. Quantum gates

A quantum gate is given by a unitary operator acting on some qubits. Such gates have matrix representations in the computational basis.

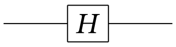
(i) The *Hadamard gate* is

$$H = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

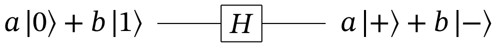
One can show that

$$H\,|0\rangle = |+\rangle; \quad H\,|1\rangle = |-\rangle; \quad H\,|+\rangle = |0\rangle; \quad H\,|-\rangle = |1\rangle$$

Note that $H^\intercal = H^\dagger = H$ and $H^2 = I$. As an orthogonal transformation in $\mathbb{R}^2$, it acts as a reflection by an angle of $\frac{\pi}{8}$ to the positive x axis. This gate is drawn



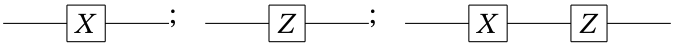
In general, by linearity we obtain



(ii) The X, Z gates are given by

$$X\,|k\rangle = |k \oplus 1\rangle\,; \quad Z\,|k\rangle = (-1)^k\,|k\rangle$$

where $\oplus$ denotes addition modulo 2. They X, Z, Y gates are drawn



(iii) The *phase gate* is

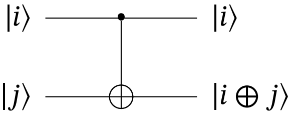
$$P_\theta = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\theta} \end{pmatrix}$$

Note that $Z = P_\pi$.

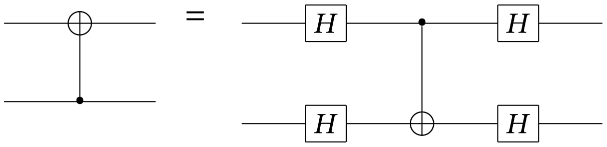
(iv) The *controlled-X* gate, also called a *CNOT* gate, is

$$CX = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & X \end{pmatrix}$$

Note that $CX\,|i\rangle\,|j\rangle = |i\rangle\,|i \oplus j\rangle$. The first qubit is called the *control* qubit, and the second is called the *target* qubit. If $i = 0$, there is no action on the second qubit. If $i = 1$, X is performed on the second qubit. In general, $CX\,|0\rangle\,|\psi\rangle = |0\rangle\,|\psi\rangle$, and $CX\,|1\rangle\,|\psi\rangle = |1\rangle\,(X\,|\psi\rangle)$. The circuit diagram is as follows.



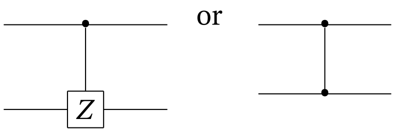
One can show that



(v) The *controlled-Z* gate, also called a *CZ* gate, is

$$CZ = \begin{pmatrix} I & 0 \\ 0 & Z \end{pmatrix}$$

So $CZ\,|0\rangle\,|\psi\rangle = |0\rangle\,|\psi\rangle$ and $CZ\,|1\rangle\,|\psi\rangle = |1\rangle\,(Z\,|\psi\rangle)$. CZ is symmetric in its action on the two qubits; for example, $CZ_{12}\,|0\rangle\,|1\rangle = CZ_{21}\,|0\rangle\,|1\rangle$. This gate is drawn



2.8. Quantum teleportation

Suppose Alice and Bob share the Bell state $|\phi^+\rangle_{AB}$, and that Alice wants to send the state of qubit $|\psi\rangle_C$ to Bob, but only classical communication between them is possible. It is possible to transfer the information about the state of $|\psi\rangle_C$ without physically transferring qubit C to Bob. This state transfer can be accomplished in such a way that is unaffected by any physical process in the space between Alice and Bob, since it relies only on classical communication.

The initial state of CAB is $|\Psi\rangle = |\psi\rangle_C \otimes |\phi^+\rangle_{AB}$, assuming $|\psi\rangle_C$ is uncorrelated with $|\phi^+\rangle_{AB}$. Let $|\psi\rangle_C = a\,|0\rangle_C + b\,|1\rangle_C$, so

$$|\Psi\rangle = |\psi\rangle_C \otimes |\phi^+\rangle_{AB} = \frac{1}{\sqrt{2}}[a\,|000\rangle + a\,|011\rangle + b\,|100\rangle + b\,|111\rangle]$$

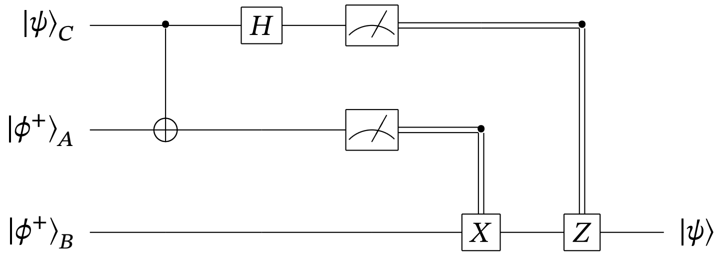
Alice sends C and A through a CX gate. Now,

$$|\Psi\rangle = |\varphi_1\rangle = \frac{1}{\sqrt{2}}[a\,|000\rangle + a\,|011\rangle + b\,|110\rangle + b\,|101\rangle]$$

She now sends C through a Hadamard gate.

$$\begin{aligned}|\Psi\rangle = |\varphi_2\rangle &= \frac{1}{\sqrt{2}}[a\,|{+}00\rangle + a\,|{+}11\rangle + b\,|{-}10\rangle + b\,|{-}01\rangle] \\ &= \frac{1}{2}\big[\,|00\rangle\,|\psi\rangle + |01\rangle\,(X\,|\psi\rangle) + |10\rangle\,(Z\,|\psi\rangle) + |11\rangle\,(-Y\,|\psi\rangle)\big]\end{aligned}$$

Alice now measures CA in the computational basis of $\mathbb{C}^2 \otimes \mathbb{C}^2$. The probability of each outcome is $\frac{1}{4}$, irrespective of the values of a and b and hence of $|\psi\rangle$. She then sends the result of her measurement to Bob. If Alice measures outcome ij, B is in state $X^j Z^i\,|\psi\rangle$. Then, Bob can act on B using $Z^i X^j$, as X and Z are involutive, giving $|\psi\rangle$ as desired. This process can be represented with the following diagram, where double-struck wires are classical, and the meter symbol denotes a measurement of the quantum state.



Note that after the measurement of CA, the entanglement between CA and B is broken. No-cloning is not violated, as the original state $|\psi\rangle_C$ is destroyed.

Note that the first steps of this process including Alice's measurement correspond to performing a Bell measurement on CA. This is because the action of CX_{CA} then H_C corresponds to a rotation of the Bell basis to the standard basis.

3. Quantum cryptography

3.1. One-time pads

We can use quantum information theory to securely transmit messages between agents Alice and Bob, who may be in distant locations, without the possibility that an eavesdropper Eve can recover the message that was sent.

We will assume that Alice and Bob have an authenticated classical channel through which they can send classical information; Alice and Bob can verify that any particular message on the channel came from a particular sender. We also assume that Eve cannot block the channel or modify any messages transmitted, but she can monitor the channel freely. Hence, Alice and Bob can receive messages from each other without error.

In the classical setting, there exists a provably secure classical scheme for private communications, called the *one-time pad*. This requires that Alice and Bob share a private key K, which is a binary string. K must have been created beforehand, and must be chosen uniformly at random from the set of binary strings of the same length as the message M. Suppose $M, K \in \{0,1\}^n$.

The protocol is as follows. First, Alice computes the encrypted message $C = M \oplus K$. She then sends C to Bob through the classical channel. Bob can then compute $C \oplus K = M \oplus K \oplus K = M$ to obtain the message that was sent by Alice. Eve cannot learn any information about the message (apart from its length), as she has no knowledge of K. In general, the probability that a particular K was chosen is 2^{-n}. This scheme cannot be broken.

Suppose that Alice and Bob use the same key K to send two messages M_1, M_2. Eve can obtain $M_1 \oplus K$ and $M_2 \oplus K$, and can therefore compute $(M_1 \oplus K) \oplus (M_2 \oplus K) = M_1 \oplus M_2$, which gives some information about the messages that were sent. Any key must only be used once, so the one-time pad protocol is inefficient. To solve this problem, we will construct methods for distributing keys, using techniques from quantum information theory.

3.2. The BB84 protocol

Quantum key distribution allows Alice and Bob to generate a private key without needing to physically meet. This key can then be used to send messages over the one-time pad protocol. In addition to a classical channel, we assume that Alice and Bob also have access to a quantum channel through which they can send qubits. We will show that Eve cannot gain information about the key that Alice and Bob generate without being detected.

Consider the bases $\mathcal{B}_0 = \{|0\rangle, |1\rangle\}, \mathcal{B}_1 = \{|+\rangle, |-\rangle\}$. These are examples of *mutually unbiased bases*; a pair of bases such that if any basis vector is measured relative to the other basis, all outcomes are equally likely. For example, measuring $|+\rangle$ relative to $\mathcal{B}_0$ gives probability $\frac{1}{2}$ for outcomes 0 and 1.

First, Alice generates two m-bit strings $x = x_1 \dots x_m \in \{0,1\}^m, y = y_1 \dots y_m \in \{0,1\}^m$ uniformly at random. She then prepares the m-qubit state $|\psi_{xy}\rangle = |\psi_{x_1 y_1}\rangle \otimes \cdots \otimes |\psi_{x_m y_m}\rangle$

where

$$|\psi_{x_i y_i}\rangle = \begin{cases} |0\rangle & x_i = 0; y_i = 0 \\ |1\rangle & x_i = 1; y_i = 0 \\ |+\rangle & x_i = 0; y_i = 1 \\ |-\rangle & x_i = 1; y_i = 1 \end{cases}$$

Alice sends the qubits $|\psi_{xy}\rangle$ to Bob with m uses of the quantum channel. The qubits received are not necessarily in the state $|\psi_{xy}\rangle$ due to noise or malicious manipulation of the channel. Bob then generates an m-bit string $y' = y'_1 \dots y'_m \in \{0,1\}^m$ uniformly at random. If $y'_i = 0$, he measures the ith qubit in the basis $\mathcal{B}_0 = \{|0\rangle, |1\rangle\}$. If $y'_i = 1$, he acts on the ith qubit by the Hadamard gate and then measures in $\mathcal{B}_0$. Equivalently, he measures the ith qubit in the basis $\mathcal{B}_1 = \{|+\rangle, |-\rangle\}$. Let the sequence of outcomes be $x' = x'_1 \dots x'_m \in \{0,1\}^m$.

If $y'_i = y_i$, we have $x'_i = x_i$. Indeed, suppose $y'_i = 0 = y_i$. Then $|\pi_{x_i y_i}\rangle \in \mathcal{B}_0$, and Bob measures in basis $\mathcal{B}_0$, so he can determine x_i with probability 1. If $y'_i = 1 = y_i$, $|\pi_{x_i y_i}\rangle \in \mathcal{B}_1$, and Bob measures in basis $\mathcal{B}_1$.

Now, Alice and Bob compare their values of y and y' over the classical channel, and discard all x_i and x'_i for which $y_i \neq y'_i$. The remaining x_i and x'_i match, given that Bob receives $|\psi_{xy}\rangle$ exactly, and this forms the shared private key $\tilde{x} = \tilde{x}'$. The average length of $\tilde{x}$ is $\frac{m}{2}$.

In the case $m = 8$, suppose $x = 01110100$ and $y = 11010001$. Alice prepares $|\psi_{xy}\rangle$ and sends the qubits to Bob. Suppose that Bob receives $|\psi_{xy}\rangle$ exactly, and he generates $y' = 01110110$. Bob measures qubit 1 in the basis $\mathcal{B}_0$, but the qubit is in state $|+\rangle$, so he obtains both outcomes for x'_1 with equal probability. He measures qubit 2 in the basis $\mathcal{B}_1$, and the qubit is in state $|-\rangle$, so after applying H and measuring, he obtains the correct outcome $x'_2 = 1$ with probability 1. After discarding mismatched y_i, the obtained private key is $\tilde{x} = 110$.

In the general case, however, there may be noise or malicious activity on the channel. We therefore include the further step of *information reconciliation* at the end of the BB84 protocol. Alice and Bob want to estimate the *bit error rate*, which is the proportion of bits in $\tilde{x}$ and $\tilde{x}'$ that differ. They can publicly compare a random sample of their bits, and discard the bits used in the test. They assume that the bit error rate in the sample is approximately the same as the bit error rate of $\tilde{x}$ and $\tilde{x}'$.

Suppose that Alice and Bob have estimated the bit error rate to be $\frac{1}{7}$, and now have strings a, b of length 7. They can use classical error correcting code techniques to fix any remaining errors. They publicly agree to act on a, b by a matrix

$$\widetilde{H} = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

which is the check matrix of a Hamming code. Alice computes the *syndrome* for a, given by $s^A = (s^A_1, s^A_2, s^A_3)^\top = \widetilde{H} a^\top$, and sends this to Bob on the public channel. Bob computes the syndrome s^B for b, and calculates $s = s^B - s^A$. There is a unique bit string v with at most one nonzero entry such that $\widetilde{H} v^\top = s$; he can therefore recover a.

The estimation of the bit error rate and the transmission of the syndrome can reveal some information on the public channel. Alice and Bob want to estimate the maximum amount of information that an eavesdropper could gain about the remaining bits, using *privacy amplification*. This depends on the choice of action that Eve takes.

As an example, suppose $a^\star = (a_1, a_2, a_3) \in \{0,1\}^3$, and suppose Eve knows at most one bit of this string. Let $c = (a_1 \oplus a_3, a_2 \oplus a_3)$. We claim that Eve has no knowledge about c. Indeed, we can explicitly enumerate all possibilities of $a^\star$ and the corresponding values of c, and show that Eve's knowledge about any of the bits of $a^\star$ does not change the distribution of c.

One strategy for Eve, called the *intercept and resend* strategy, is to intercept the qubits as they are transferred to Bob, measure them, and retransmit the post-measurement state. The best possible measurement she can perform is in the *Breidbart basis* $\{\lvert\alpha_0\rangle, \lvert\alpha_1\rangle\}$ where

$$\lvert\alpha_0\rangle = \cos\frac{\pi}{8}\lvert 0\rangle - \sin\frac{\pi}{8}\lvert 1\rangle; \quad \lvert\alpha_1\rangle = \sin\frac{\pi}{8}\lvert 0\rangle + \cos\frac{\pi}{8}\lvert 1\rangle$$

Note that

$$\lvert\langle\alpha_0|0\rangle\rvert^2 = \lvert\langle\alpha_0|+\rangle\rvert^2 = \cos^2\frac{\pi}{8}; \quad \lvert\langle\alpha_1|1\rangle\rvert^2 = \lvert\langle\alpha_1|-\rangle\rvert^2 = \cos^2\frac{\pi}{8}$$

The $\lvert\alpha_i\rangle$ provide the best possible simultaneous approximations of $\lvert 0\rangle, \lvert +\rangle$ and $\lvert 1\rangle, \lvert -\rangle$. Suppose $y_i' = y_i$, and suppose Eve intercepts the ith qubit and measures it in the Breidbart basis. Her outcomes are 0 or 1, and she learns the correct value of x_i with probability $\cos^2\frac{\pi}{8} \approx 0.85$. If she measures 0, she transmits $\lvert\alpha_0\rangle$ to Bob, and if she measures 1, she transmits $\lvert\alpha_1\rangle$ to Bob.

The probability that Bob makes an incorrect inference of the value of the ith bit after this manipulation is $\frac{1}{4}$, regardless of the state of the qubit transmitted by Alice. Suppose $\lvert\psi_{x_i y_i}\rangle = \lvert 0\rangle$, so $x_i = 0, y_i = 0$. Then,

$$\begin{aligned}
\mathbb{P}\left(x_i' \neq x_i\right) &= \mathbb{P}\left(B \text{ measures } 1 \mid A \text{ sent } \lvert 0\rangle\right) \\
&= \mathbb{P}\left(E \text{ sent } \lvert\alpha_0\rangle \mid A \text{ sent } \lvert 0\rangle\right)\mathbb{P}\left(B \text{ measures } 1 \mid E \text{ sent } \lvert\alpha_0\rangle\right) \\
&+ \mathbb{P}\left(E \text{ sent } \lvert\alpha_1\rangle \mid A \text{ sent } \lvert 0\rangle\right)\mathbb{P}\left(B \text{ measures } 1 \mid E \text{ sent } \lvert\alpha_1\rangle\right) \\
&= \lvert\langle\alpha_0|0\rangle\rvert^2\lvert\langle\alpha_0|1\rangle\rvert^2 + \lvert\langle\alpha_1|0\rangle\rvert^2\lvert\langle\alpha_1|1\rangle\rvert^2 \\
&= \frac{1}{4}
\end{aligned}$$

4. Quantum computation

4.1. Classical computation

A *computational task* takes an input bit string and produces an output bit string.

A decision problem is a computational task that produces an output of length 1. Let $B = B_1 = \{0, 1\}$ and denote $B_n = \{0, 1\}^n$. Define $B^\star = \bigcup_{n \geq 1} B_n$. A *language* is a subset $L \subseteq B^\star$. A decision problem corresponds to the problem of checking whether a word $w \in B^\star$ lies in a language L. For example, the set of primes, expressed in binary, forms a language $P \subseteq B^\star$, and there is a corresponding decision problem to check if a given binary string represents a prime.

More generally, the output of a computational task can be of any length. For example, the task $\mathsf{FACTOR}(x)$ takes the input x and produces a bit string containing a factor of x, or 1 if x is prime.

There are various models of computation, but we restrict to the *circuit* or *gate array* model. In this model, we have an input $x = b_1 \dots b_n \in B_n$, and extend it with some trailing zeroes to add scratch space to perform computations. We then perform some computational steps, an application of designated Boolean gates $f : B_n \to B_m$ on preassigned bits. For each n, we have a circuit C_n, which is a prescribed sequence of computational steps that performs a given task for all inputs of size n. The output to the computation is a designated subsequence of the extended bit string.

Suppose that, in addition to extending the input bit string with zeroes, we also add k random bits, which have values set to 0 or 1 uniformly at random. The output of the computation will now be probabilistic. The probability that the output is y is $a2^{-k}$, where a is the number of bit strings r that produce the desired outcome. We typically require that the output is correct with some prescribed probability.

4.2. Classical complexity

The *time complexity* is a measure of the amount of computational steps required for a particular algorithm for an input of size n. In the circuit model, we define $T(n)$ to be the total number of gates in the circuit C_n, known as the *size* of the circuit or *runtime* of the algorithm.

For a positive function $T(n)$, we write $T(n) = O(f(n))$ if there exist positive constants c, n_0 such that for all $n > n_0$, we have $T(n) \leq cf(n)$. If $T(n) = O(n^k)$ for some $k > 0$, we say that $T(n)$ is $O(\text{poly}(n))$, and the corresponding algorithm is a *poly-time* algorithm. The class of languages for which the membership problem has a classical poly-time algorithm is called P. The class of languages for which the membership problem has a randomised classical poly-time algorithm that gives the correct answer with probability at least $\frac{2}{3}$ is called BPP, short for *bounded-error probabilistic poly-time*. The problem $\mathsf{FACTOR}(M, N)$ which determines if there is a nontrivial factor of N that is at most M does not lie in BPP. The best known runtime is $T(n) = O\left(n^{\frac{1}{3}}(\log n)^{\frac{2}{3}}\right)$.

A black box promise problem is a computational task where the input is a *black box* or *oracle* which can compute a Boolean function $f : B_m \to B_n$, and there is an *a priori promise* on f restricting the possible values of f. For example, the black box promise problem for constant vs. balanced functions takes a function $f : B_n \to B$ such that f is constant or *balanced*, in which case f is equal to zero for exactly half of the 2^n possible inputs.

The corresponding complexity is called *query complexity*, which counts the amount of times we need to query the black box. We typically wish to minimise the query complexity.

4.3. Quantum circuits

In a quantum circuit, we have qubit inputs $|b_1\rangle \dots |b_n\rangle |0\rangle \dots |0\rangle$ analogously to the classical case. The input size n is the number of qubits. The addition of randomness to classical computation needs no analogue in the quantum case, since randomness is obtained by measurement. For instance, if we have a qubit $|0\rangle$, we can generate a uniform Bernoulli random variable by sending the qubit through a Hadamard gate and then measuring in the computational basis.

The computational steps are gates or unitary operators, which act on a prescribed set of qubits, constituting a quantum circuit C_n. The output is obtained by performing a measurement on a prescribed set of qubits. One can show that any circuit involving arbitrarily many measurements is equivalent to a circuit that only performs a single measurement at the end of the computation.

4.4. Quantum oracles

Note that all quantum gates are invertible, as they are represented with unitary operators, but not all classical gates are invertible. Any $f : B_m \to B_n$ can be expressed in an equivalent invertible form $\widetilde{f} : B_{m+n} \to B_{m+n}$ by defining $\widetilde{f}(b, c) = (b, c \oplus f(b))$. If we can compute f we can also compute $\widetilde{f}$, and conversely given $\widetilde{f}$ we can find $f(b) = \widetilde{f}(b, 0)$. This is self-inverse.

$$\widetilde{f}(\widetilde{f}(b, c)) = \widetilde{f}(b, c \oplus f(b)) = (b, c \oplus f(b) \oplus f(b)) = (b, c)$$

A quantum oracle for a function $f : B_m \to B_n$ is the quantum gate U_f acting on $m+n$ qubits such that $U_f |x\rangle |y\rangle = |x\rangle |y \oplus f(x)\rangle$ for $|x\rangle, |y\rangle$ states in the computational basis. In other words, its action on the computational basis is $\widetilde{f}$. We say that $|x\rangle$ is the *input register* and $|y\rangle$ is the *output register*.

One can show that U_f is always a unitary operator. We can show this directly by considering $U_f |x'\rangle |y'\rangle = |x'\rangle |y' \oplus f(x')\rangle$, and we can take the inner product with $U_f |x\rangle |y\rangle = |x\rangle |y \oplus f(x)\rangle$. An easier way to show this is to consider $\widetilde{f} : B_k \to B_k$ as a permutation on B_k where $m + n = k$. We can write $U_f |x\rangle |y\rangle = U_f |i_1 \dots i_k\rangle = \left|\widetilde{f}(i_1 \dots i_k)\right\rangle$. Since $\widetilde{f}$ is a permutation, U_f is therefore represented by a permutation matrix, which has a single 1 in each row and column. All permutation matrices are unitary.

In contrast to a classical oracle, a quantum oracle can act on a superposition of input registers. Let $f : B_m \to B_n$, and consider the *equal superposition* state $|\varphi_m\rangle = \frac{1}{\sqrt{2^m}} \sum_{x \in B_m} |x\rangle$. We can find

$$U_f |\varphi_m\rangle |y\rangle = U_f\left(\frac{1}{\sqrt{2^m}} \sum_{x \in B_m} |x\rangle\right) |y\rangle = \frac{1}{\sqrt{2^m}} \sum_{x \in B_m} U_f |x\rangle |y\rangle = |\psi_f\rangle$$

In a single use of the oracle, we obtain a final state which depends on the value of f corresponding to all possible inputs. One can easily create such an equal superposition state $|\varphi_m\rangle$ by sending the m-qubit state $|0\rangle \dots |0\rangle$ through m Hadamard gates $H \otimes \cdots \otimes H$. We have $(H|0\rangle)^{\otimes m} = (|+\rangle)^{\otimes m} = |\varphi_m\rangle$. This creates a superposition of exponentially many terms using a linear amount of Hadamard gates.

4.5. Deutsch–Jozsa algorithm

Consider the black box problem for balanced vs. constant functions. Classically, one needs $2^{n-1} + 1$ queries to solve the problem in the worst case. This amount of queries is clearly sufficient; even if f is balanced, the first 2^{n-1} queries could have equal outcomes, but the subsequent query must have a different outcome. Suppose that there exists an algorithm that can solve the problem in 2^{n-1} queries. An adversary that controls the oracle can respond with 0 for every query, and subsequently choose a function f that agrees with the earlier query results but is balanced or constant as required to cause the algorithm to produce an error. Therefore, classically we require a query complexity of $O(\exp(n))$.

Suppose we have a quantum oracle U_f with $U_f |x\rangle |y\rangle = |x\rangle |y \oplus f(x)\rangle$, where $|x\rangle$ is an n-qubit state and $|y\rangle$ is a 1-qubit state. Set each qubit to state $|0\rangle$, then act by $H^{\otimes n} \otimes (H \cdot X)$ on $|x\rangle |y\rangle$. We then obtain the state $|A\rangle = \frac{1}{\sqrt{2^n}} \sum_{x \in B_n} |x\rangle |-\rangle$. Send this state through the oracle to obtain $U_f |A\rangle = \frac{1}{\sqrt{2^n}} U_f \sum_{x \in B_n} |x\rangle |-\rangle$. Note that

$$\begin{aligned}
U_f |x\rangle |-\rangle &= \frac{1}{\sqrt{2}} U_f(|x\rangle |0\rangle - |x\rangle |1\rangle) \\
&= \frac{1}{\sqrt{2}}(|x\rangle |f(x)\rangle - |x\rangle |f(x)^c\rangle) \\
&= \begin{cases} \frac{1}{\sqrt{2}} |x\rangle (|0\rangle - |1\rangle) = |x\rangle |-\rangle & \text{if } f(x) = 0 \\ \frac{1}{\sqrt{2}} |x\rangle (|1\rangle - |0\rangle) = -|x\rangle |-\rangle & \text{if } f(x) = 1 \end{cases} \\
&= (-1)^{f(x)} |x\rangle |-\rangle
\end{aligned}$$

The method of encoding all information into a phase is called *phase kickback*. Hence,

$$U_f |A\rangle = \frac{1}{\sqrt{2^n}} U_f \sum_{x \in B_n} |x\rangle |-\rangle = \frac{1}{\sqrt{2^n}} \left(\sum_{x \in B_n} (-1)^{f(x)} |x\rangle\right) |-\rangle$$

We can then easily discard the last qubit, as it is now in a product state. We obtain

$$\lvert f\rangle = \frac{1}{\sqrt{2^n}} \sum_{x\in B_n} (-1)^{f(x)} \lvert x\rangle$$

If f is constant,

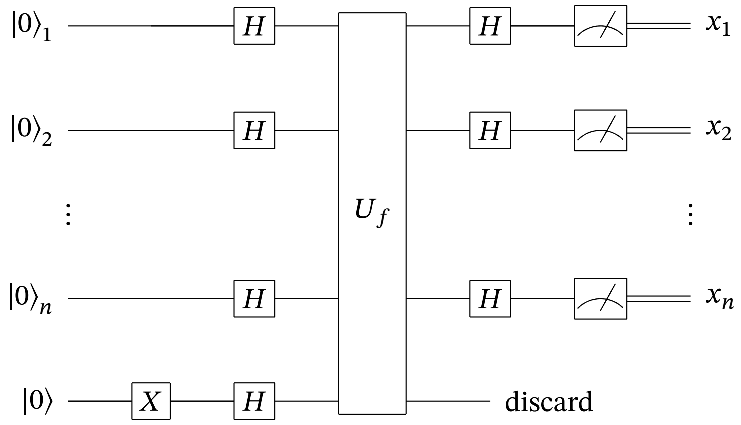
$$\lvert f\rangle = \pm\frac{1}{\sqrt{2^n}} \sum_{x\in B_n} \lvert x\rangle = \pm(H\lvert 0\rangle)^{\otimes n}$$

If we apply $H^{\otimes n}$ to $\lvert f\rangle$, we obtain $\pm\lvert 0\rangle^{\otimes n}$. If f is balanced, writing $\lvert \varphi_n\rangle = \frac{1}{\sqrt{2^n}} \sum_{y\in B_n} \lvert y\rangle$,

$$\langle f\vert\varphi_n\rangle = \frac{1}{2^n} \sum_{x,y\in B_n} (-1)^{f(x)} \langle y\vert x\rangle = \frac{1}{2^n} \sum_{x\in B_n} (-1)^{f(x)} = 0$$

In this case, $\lvert f\rangle$ is orthogonal to $\lvert\varphi_n\rangle$. Applying $H^{\otimes n}$ to $\lvert f\rangle$, we have that $H^{\otimes n}\lvert f\rangle$ is orthogonal to $H^{\otimes n}\lvert\varphi_n\rangle = \lvert 0\rangle^{\otimes n}$.

After obtaining $\lvert f\rangle$, we apply $H^{\otimes n}$ and measure in the computational basis. If f is constant, we measure $0 \dots 0$ with probability 1, and if f is balanced, we measure $0 \dots 0$ with probability 0. This allows us to infer whether f is constant or balanced with probability 1.



For this algorithm, we use one query and $3n + 2$ further operations.

Suppose we permit a probability $\varepsilon > 0$ of error. In the quantum case, we only need one query. In the classical case, there is a randomised algorithm which solves the problem with a constant number $O\left(\log\frac{1}{\varepsilon}\right)$ of queries for all n. Choose k inputs each chosen uniformly at random, and evaluate $f(x)$ for each x in this set. If $f(x)$ is constant for all of these k inputs, we infer f is constant; otherwise we infer it is balanced. An error can only occur when the function is balanced but we infer it is constant. The probability of error is $\frac{2}{2^k} = 2^{-k+1}$. Hence, we can take $\varepsilon < 2^{-k+1}$, so $k = O\left(\log\frac{1}{\varepsilon}\right)$.

4.6. Simon's algorithm

Consider a function $f : B_n \to B_n$ with the promise that either f is injective, or $f(x) = f(y)$ if and only if $y = x$ or $y = x \oplus \xi$ for a fixed $0 \neq \xi \in B_n$. The problem is to determine with bounded error whether f is in the 1-1 form or the 2-1 form, and in the latter case, to find the constant ξ. Note that $f(x \oplus \xi) = f(x)$ is the statement that f has period ξ.

Classically, the query complexity is $O(\exp(n))$. In order to solve the problem, we need to find two distinct x, y inputs for which $f(x) = f(y)$, or show that this is not possible. However, there is a quantum algorithm with query complexity $O(n)$.

4.7. Quantum Fourier transform

Let $\mathcal{V}_N$ be a state space, and $\mathcal{B}_N = \{|0\rangle, |1\rangle, \dots, |N-1\rangle\}$ be an orthonormal basis for $\mathcal{V}_N$. Write $\mathbb{Z}_N$ for integers modulo N, and let $\omega = e^{\frac{2\pi i}{N}}$. For $|k\rangle \in \mathcal{B}_N$, we define

$$QFT_N |k\rangle = \frac{1}{\sqrt{N}} \sum_{\ell=0}^{N-1} e^{\frac{2\pi i}{N} k\ell} |\ell\rangle = \frac{1}{\sqrt{N}} \sum_{\ell=0}^{N-1} \omega^{k\ell} |\ell\rangle$$

The quantum Fourier transform can be viewed as a generalisation of the Hadamard operator, as $QFT_2 = H$.

We show that this is a unitary operator.

$$(QFT)_{jk} = \langle j| QFT |k\rangle = \frac{1}{\sqrt{N}} \sum_{\ell=0}^{N-1} \omega^{k\ell} \langle j|\ell\rangle = \frac{1}{\sqrt{N}} \omega^{jk}$$

$$QFT = \frac{1}{\sqrt{N}} \begin{pmatrix} 1 & 1 & 1 & 1 & \cdots \\ 1 & \omega & \omega^2 & \omega^3 & \cdots \\ 1 & \omega^2 & \omega^4 & \omega^6 & \cdots \\ 1 & \omega^3 & \omega^6 & \omega^9 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Let S_j be the sum of the jth row or column. If $j = 0$, $S_j = \frac{1}{\sqrt{N}} N$. Otherwise,

$$S_j = \frac{1}{\sqrt{N}} (1 + \omega^j + \dots + \omega^{j(N-1)}) = \frac{1}{\sqrt{N}} \cdot \frac{1 - \omega^{jN}}{1 - \omega^j} = 0$$

We can use this to prove that $(QFT^\dagger QFT)_{jk} = \delta_{jk}$, so it is a unitary operator.

Suppose we have a periodic function $f : \mathbb{Z}_N \to Y$, where typically $Y = \mathbb{Z}_M$ for some M. Let r be the smallest integer in $\mathbb{Z}_N$ for which $f(x + r) = f(x)$ for all $x \in \mathbb{Z}_N$, so f is periodic with period r. Suppose further that f is injective in each period. We wish to find r with a particular probability of error.

There is a classical algorithm with query complexity $O(\sqrt{N}) = O\left(2^{\log N^{\frac{1}{2}}}\right) = O\left(2^{\frac{1}{2}\log N}\right)$. In the quantum case, for any error probability $\varepsilon \in (0,1)$, there is an algorithm with query complexity $O(\log\log N)$, which provides an exponential speed increase.

We first describe an attempt to construct such an algorithm without using the quantum Fourier transform. Begin with the uniform superposition state $|\psi_N\rangle = \frac{1}{\sqrt{N}}\sum_{x=0}^{N-1}|x\rangle$. Consider the quantum oracle U_f corresponding to $f : \mathbb{Z}_N \to \mathbb{Z}_M$, defined by $U_f|x\rangle|y\rangle = |x\rangle|y + f(x)\rangle$, where addition is performed modulo M. Set the output register $|y\rangle$ to $|0\rangle$, and then compute $|f\rangle = U_f|\psi_N\rangle|0\rangle$. We obtain

$$|f\rangle = U_f|\psi_N\rangle|0\rangle = \frac{1}{\sqrt{N}}\sum_{x=0}^{N-1} U_f|x\rangle|0\rangle = \frac{1}{\sqrt{N}}\sum_{x=0}^{N-1}|x\rangle|f(x)\rangle$$

Since r is the period, we have $r \mid N$, so let $A = \frac{N}{r} \in \mathbb{N}$ be the number of periods. We now measure the second register, giving an outcome $y = f(x_0)$ for some $x_0 \in \{0,\dots,r-1\}$. Note that $y = f(x_0 + jr)$ for any $j \in \{0,\dots,A-1\}$. The terms in $|f\rangle$ which contribute to the outcome $y = f(x_0)$ are

$$\frac{1}{\sqrt{N}}\sum_{j=0}^{A-1}|x_0 + jr\rangle|f(x_0)\rangle$$

Hence, the probability of obtaining a particular outcome $f(x_0)$ is $\frac{A}{N} = \frac{1}{r}$. Then, the post-measurement state of the input register is

$$|\mathrm{per}\rangle = \frac{1}{\sqrt{A}}\sum_{j=0}^{A-1}|x_0 + jr\rangle$$

The state $|\mathrm{per}\rangle$ is periodic. If we measure the input register, we obtain $|x_0 + j_0 r\rangle$ for some $j_0 \in \{0,\dots,A-1\}$, selected uniformly at random. The probability that the outcome of this second measurement is $x_0 + j_0 r$ is $\frac{1}{A}$. Therefore, no information about r is obtained.

We resolve this issue by utilising the quantum Fourier transform. Instead of measuring the input register, we act on $|\mathrm{per}\rangle$ by QFT_N. Since

$$QFT_N|x\rangle = \frac{1}{\sqrt{N}}\sum_{y=0}^{N-1}\omega^{xy}|y\rangle$$

we find

$$\begin{aligned}
QFT_N \left|\text{per}\right\rangle &= \frac{1}{\sqrt{A}} \sum_{y=0}^{N-1} QFT_N \left|x_0 + jr\right\rangle \\
&= \frac{1}{\sqrt{A}} \frac{1}{\sqrt{N}} \sum_{j=0}^{A-1} \sum_{y=0}^{N-1} \omega^{(x_0+jr)y} \left|y\right\rangle \\
&= \frac{1}{\sqrt{NA}} \sum_{y=0}^{N-1} \omega^{x_0 y} \underbrace{\left[\sum_{j=0}^{A-1} (\omega^{ry})^j \right]}_{S} \left|y\right\rangle
\end{aligned}$$

Note that

$$S = \begin{cases} A & \text{if } \omega^{ry} = 1 \\ \frac{1-\omega^{ryA}}{1-\omega^{ry}} = 0 & \text{otherwise} \end{cases}$$

Note that $\omega^{ry} = 1$ if $y = kA = \frac{kN}{r}$ for $k \in \{0, \dots, r-1\}$. Hence, we obtain

$$QFT_N \left|\text{per}\right\rangle = \frac{A}{\sqrt{NA}} \sum_{k=0}^{r-1} \omega^{x_0 \frac{kN}{r}} \left|\frac{kN}{r}\right\rangle = \frac{1}{\sqrt{r}} \sum_{k=0}^{r-1} \omega^{x_0 \frac{kN}{r}} \left|\frac{kN}{r}\right\rangle$$

The value of x_0 is no longer present in a ket, and has been converted into phase information. It therefore does not affect measurement outcomes. The periodicity in r has been inverted into periodicity in $\frac{1}{r}$. The resulting state is still periodic, but each period begins at 0 instead of x_0.

Now, when measuring this register, the outcome is $c = \frac{k_0 N}{r}$ for some $k_0 \in \{0, \dots, r-1\}$. Each outcome occurs with probability r. Note that $\frac{k_0}{r} = \frac{c}{N}$, and $\frac{c}{N}$ is known after performing the measurement; we wish to know the value of r.

Suppose first that k_0 is coprime to r. In this case, we can cancel $\frac{c}{N}$ to its lowest form, then the denominator is r. If k_0 is not coprime to r, the denominator $\tilde{r}$ will instead be a factor of r. To solve this, we can compute the reduced denominator and then evaluate $f(0), f(\tilde{r})$; if they are equal, $\tilde{r} = r$, and otherwise, $\tilde{r} \mid r$. We would like to know the probability that a randomly chosen k_0 is coprime to the true periodicity r.

Theorem (coprimality theorem). Let $\varphi(r)$ denote the number of integers less than r that are coprime to r. Then there exist $c > 0, r_0 > 0$ such that for all $r \geq r_0$, $\varphi(r) \geq c\frac{r}{\log\log r}$. In particular, $\varphi(r) = \Omega\left(\frac{r}{\log\log r}\right)$.

This theorem implies that since k_0 is chosen uniformly at random, the probability that k_0 is coprime to r is $O\left(\frac{1}{\log\log r}\right)$. We claim that if we repeat this process $O(\log\log r)$ times, we will obtain an outcome c such that after cancellation, $\frac{c}{N} = \frac{k_0}{r}$ where k_0 is coprime to r in at least one case, with a constant probability. This claim follows from the following lemma.

Lemma. Suppose that a single trial has success probability p, and the trial is repeated M times independently, for any $\varepsilon \in (0,1)$, the probability of at least one success is greater than $1-\varepsilon$ if $M = \frac{-\log\varepsilon}{p}$.

Therefore, to achieve a constant probability $1-\varepsilon$ of success, we need $O\left(\frac{1}{p}\right)$ trials. In the algorithm above, $p = O\left(\frac{1}{\log\log r}\right)$, so we need $O(p) = O(\log\log r) < O(\log\log N)$ trials to achieve the desired result.

In each invocation of the algorithm, we query f three times: once to construct the state $|f\rangle$, and twice to check if $\tilde{r}$ is the true periodicity. We also need to apply the quantum Fourier transform QFT_N, which has implementations in $O((\log N)^2)$ steps. We must also perform standard arithmetic operations such as to cancel denominators, which are computable in $O(\text{poly}(\log N))$ steps. Therefore, we succeed in determining the period with any constant probability of success $1-\varepsilon$ with $O(\log\log N)$ queries and $O(\text{poly}(\log N))$ additional steps.

4.8. Efficient implementation of quantum Fourier transform

We can implement a quantum Fourier transform using $O(\text{poly}(\log N))$ gates if $N = 2^n$. In this case, QFT_N acts on n qubits. If $N \neq 2^n$, we do not have an efficient implementation; in this case, we approximate N by 2^k for some $k \in \mathbb{Z}$. In the case $N = 2^n$, we demonstrate a quantum circuit of size $O(n^2)$.

If $x \in \mathbb{Z}_n = \{0, \dots, 2^n - 1\}$, note that

$$QFT_N\,|x\rangle = \frac{1}{\sqrt{N}} \sum_{y=0}^{N-1} \omega^{xy}\,|y\rangle$$

We can represent x and y by n-bit strings.

$$x = (x_0, x_1, \dots, x_{n-1}); \quad x = \sum_{i=0}^{n-1} 2^i x_i$$

Now, $\omega^{xy} = \exp\left[\frac{2\pi i}{2^n} xy\right]$.

$$\frac{xy}{2^n} = \frac{1}{2^n}\left[(x_0 + 2x_1 + \dots + 2^{n-1}x_{n-1})(y_0 + 2y_1 + \dots + 2^{n-1}y_{n-1})\right]$$

Retaining only the fractional terms of $\frac{xy}{2^n}$, as integral parts do not contribute to the final result, we obtain

$$y_{n-1}(.x_0) + y_{n-2}(.x_1x_0) + \dots + y_0(.x_{n-1}\dots x_0)$$

where for instance $.x_1x_0 = \frac{x_1}{2} + \frac{x_0}{2^2}$. Hence,

$$
\begin{aligned}
QFT\,|x\rangle &= \frac{1}{\sqrt{2^n}} \sum_{y_0,\dots,y_{n-1}\in\{0,1\}} \exp\left[\frac{2\pi i x y}{2^n}\right] |y_{n-1}\rangle \dots |y_0\rangle \\
&= \left(\frac{1}{\sqrt{2}} \sum_{y_{n-1}\in\{0,1\}} \exp\left[2\pi i y_{n-1}(.x_0)\right] |y_{n-1}\rangle\right) \cdots \left(\frac{1}{\sqrt{2}} \sum_{y_0\in\{0,1\}} \exp\left[2\pi i y_0(.x_{n-1}\dots x_0)\right] |y_0\rangle\right) \\
&= \frac{1}{\sqrt{2}}\big(|0\rangle + e^{2\pi i(.x_0)}\,|1\rangle\big) \dots \frac{1}{\sqrt{2}}\big(|0\rangle + e^{2\pi i(.x_{n-1}\dots x_0)}\,|1\rangle\big)
\end{aligned}
$$

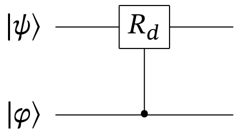
To implement the quantum Fourier transform, we will use the Hadamard gate, the 1-qubit phase gate, and the 2-qubit controlled phase gate. Note that we can write

$$H\,|x\rangle = \frac{1}{\sqrt{2}}\big[|0\rangle + e^{2\pi i(.x)}\,|1\rangle\big]$$

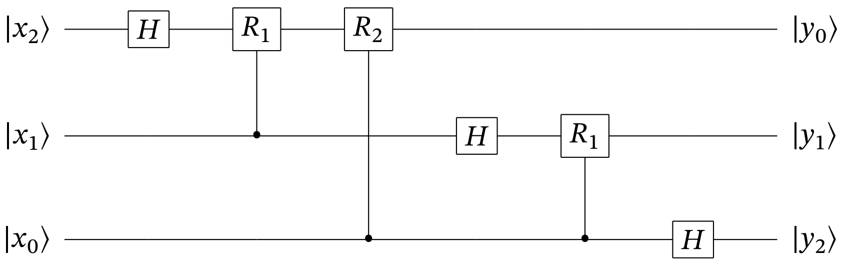
For any $d \in \mathbb{Z}_+$, the phase gate is given by

$$R_d = \begin{pmatrix} 1 & 0 \\ 0 & \exp\left[\frac{i\pi}{2^d}\right] \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \exp\left[2\pi i(.\underbrace{0\dots0}_{d\text{ zeroes}}1)\right] \end{pmatrix}$$

Note that $R_d\,|0\rangle = |0\rangle$ and $R_d\,|1\rangle = e^{2\pi i(.0\dots01)}\,|1\rangle$. In the case $d = 1$, we obtain $R_1\,|1\rangle = e^{2\pi i(.01)}\,|1\rangle = i\,|1\rangle$. The two-qubit controlled phase gate, denoted CR_d, is drawn



If $|\varphi\rangle = |0\rangle$, $CR_d\,|0\rangle\,|\psi\rangle = |0\rangle\,|\psi\rangle$. If $|\varphi\rangle = |1\rangle$, $CR_d\,|1\rangle\,|\psi\rangle = |1\rangle\,R_d\,|\psi\rangle$. We will now describe the quantum circuit for QFT_8, so $N = 8$ and $n = 3$.



Applying the given gates to $|x_2\rangle$, we obtain

$$\begin{aligned}
|x_2\rangle &\xrightarrow{H} \frac{1}{\sqrt{2}}\big[|0\rangle + e^{2\pi i(.x_2)}\,|1\rangle\big] \\
&\xrightarrow{R_1} \frac{1}{\sqrt{2}}\big[|0\rangle + e^{2\pi i(.x_2)}e^{2\pi i(.0x_1)}\,|1\rangle\big] \\
&\xrightarrow{R_2} \frac{1}{\sqrt{2}}\big[|0\rangle + e^{2\pi i(.x_2)}e^{2\pi i(.0x_1)}e^{2\pi i(.00x_0)}\,|1\rangle\big] \\
&= \frac{1}{\sqrt{2}}\big[|0\rangle + e^{2\pi i(.x_2x_1x_0)}\,|1\rangle\big] = |y_0\rangle
\end{aligned}$$

as required. Typically, after applying the above circuit, we will swap the states $|y_0\rangle, |y_1\rangle, |y_2\rangle$ to be in reverse order; this takes $O(n)$ gates.

In this implementation, we used 3 Hadamard gates, and $2 + 1 = 3$ controlled phase gates. If $N = 2^n$, we need n Hadamard gates and $\frac{n(n-1)}{2} = O(n^2)$ controlled phase gates.

4.9. Grover's algorithm

Suppose we have a large unstructured database of N items, in which we aim to locate a particular 'good' item. Suppose that given an item, we can easily check if it is the 'good' item. We wish to construct an algorithm to locate this good item with success probability at least $1 - \varepsilon$. Each access to the database is considered a query.

In the classical case, we need $O(N)$ queries: if we find a bad item, it gives us no information about the location of the good item. The probability that any item is good is $\frac{1}{N}$. Given M queries, the probability of success is $\frac{M}{N} \geq 1 - \varepsilon$, so $M \geq (1-\varepsilon)N$ gives $M = O(N)$. In the quantum case, $O(\sqrt{N})$ queries are necessary and sufficient. This is not an exponential speedup but a quadratic speedup.

Let $\mathcal{V}$ be a vector space, and let $|v\rangle \in \mathcal{V}$. We define the rank 1 projection operator $\Pi_{|\alpha\rangle} = |\alpha\rangle\langle\alpha|$, and the reflection operator $I_{|\alpha\rangle} = I - 2\,|\alpha\rangle\langle\alpha|$. Note that $I_{|\alpha\rangle}\,|\alpha\rangle = -\,|\alpha\rangle$. Let $|\psi\rangle \in \mathcal{S}^{\perp}_{|v\rangle} = \operatorname{span}\{|\beta\rangle \in \mathcal{V} \mid \langle\alpha|\beta\rangle = 0\}$. Then $I_{|\alpha\rangle}\,|\psi\rangle = |\psi\rangle - |\alpha\rangle\langle\alpha|\psi\rangle = |\psi\rangle$.

For any unitary operator U acting on $\mathcal{V}$, we have $U\Pi_{|\alpha\rangle}U^\dagger = U\,|\alpha\rangle\langle\alpha|\,U^\dagger = \Pi_{U|\alpha\rangle}$. Note also that $UI_{|\alpha\rangle}U^\dagger = U(I - 2\,|\alpha\rangle\langle\alpha|)U^\dagger = I - 2\,|U\alpha\rangle\langle U\alpha| = I_{U|\alpha\rangle}$.

If $\mathcal{V} = \mathbb{C}^2$, for all $|\alpha\rangle \in \mathcal{V}$, let $|\alpha^\perp\rangle$ be orthogonal to $|\alpha\rangle$. For all $|v\rangle \in \mathcal{V}$, we can write $|v\rangle = a\,|\alpha\rangle + b\,|\alpha^\perp\rangle$, so $\Pi_{|\alpha\rangle}\,|v\rangle = a\,|\alpha\rangle$ and $I_{|\alpha\rangle}\,|v\rangle = -a\,|\alpha\rangle + b\,|\alpha^\perp\rangle$.

Let $N = 2^n$, so we can label each item in the database with an n-bit binary string. We will convert the search problem into a black-box promise problem. The database corresponds to the Boolean function $f : B_n \to B$ where $f(x_0) = 1$ for a particular $x_0 \in B_n$, and $f(x) = 0$ otherwise. The corresponding quantum oracle is $U_f\,|x\rangle\,|y\rangle = |x\rangle\,|y \oplus f(x)\rangle$, where $|x\rangle \in$

$(\mathbb{C}^2)^{\otimes n}$ and $|y\rangle \in \mathbb{C}^2$. The fact that the database is unstructured corresponds to the fact that the quantum oracle U_f is a black box. We will use the operator I_{x_0}, which has the following action on the basis vectors.

$$I_{x_0}|x\rangle = \begin{cases} +|x\rangle & \text{if } x \neq x_0 \\ -|x\rangle & \text{if } x = x_0 \end{cases}$$

If $x_0 = 0 \dots 0 \in B_n$, we define $I_0 = I_{x_0}$. Note that I_{x_0} can be implemented using U_f; indeed,

$$\begin{aligned} U_f|x\rangle|-\rangle &= \frac{1}{\sqrt{2}}U_f|x\rangle(|0\rangle - |1\rangle) \\ &= \frac{1}{\sqrt{2}}(|x\rangle|f(x)\rangle - |x\rangle|f(x)^c\rangle) \\ &= \begin{cases} \frac{1}{\sqrt{2}}|x\rangle(|0\rangle - |1\rangle) & \text{if } x \neq x_0 \\ \frac{1}{\sqrt{2}}|x\rangle(|1\rangle - |0\rangle) & \text{if } x = x_0 \end{cases} \\ &= \begin{cases} +|x\rangle|-\rangle & \text{if } x \neq x_0 \\ -|x\rangle|-\rangle & \text{if } x = x_0 \end{cases} \end{aligned}$$

Hence, $U_f|x\rangle|-\rangle = (I_{x_0}|x\rangle)|-\rangle$. So if $|\psi\rangle \in (\mathbb{C}^2)^{\otimes n}$, $|\psi\rangle = a_0|x_0\rangle + \sum_{x\neq x_0} a_x|x\rangle$ gives $U_f|\psi\rangle|-\rangle = (I_{x_0}|\psi\rangle)|-\rangle = -a_0|x_0\rangle + \sum_{x\neq x_0} a_x|x\rangle$.

Given a black box which computes I_{x_0} for some $x_0 \in B_n$, we wish to determine x_0 with the least amount of queries. We will now describe Grover's algorithm. We begin with the equal superposition state $|\psi_0\rangle = \frac{1}{\sqrt{2^n}}\sum_{x\in B_n}|x\rangle$. Consider *Grover's iteration operator* $Q = -H_n I_0 H_n I_{x_0}$ where $H_n = H^{\otimes n}$. Note that Q is real-valued, so acts geometrically on the real-valued vector $|\psi_0\rangle$ in real Euclidean space. It has the following properties.

(i) In the plane $\mathcal{P}(x_0)$ spanned by $|x_0\rangle$ and $|\psi_0\rangle$, Q acts as a rotation through an angle 2α where $\sin\alpha = \frac{1}{\sqrt{2^n}}$.

(ii) In the plane orthogonal to $\mathcal{P}(x_0)$, Q acts as $-I$.

We repeatedly apply Q to $|\psi_0\rangle$ to obtain the rotated vector $|\psi_0'\rangle$, and then measure in the computational basis.

$$|\psi_0'\rangle = a_0|x_0\rangle + \sum_{x_i\neq x_0}\sum a_i|x_i\rangle$$

Hence, the probability that the outcome is x_0 is $|a_0|^2 = |\langle x_0|\psi_0'\rangle|^2 = |\cos\delta|^2 \approx 1$ where δ is the angle between $|\psi_0'\rangle$ and $|x_0\rangle$.

If n is large, $|\psi_0\rangle$ is almost orthogonal to $|x_0\rangle$, with $\langle x_0|\psi_0\rangle = \frac{1}{\sqrt{2^n}} = \cos\beta$. By property (i), Q acting on $|\psi_0\rangle$ rotates the state by 2α, where $\sin\alpha = \frac{1}{\sqrt{2^n}}$. Let m be the number of iterations needed to rotate $|\psi_0\rangle$ close to $|x_0\rangle$. Then

$$m = \frac{\beta}{2\alpha} = \frac{\arccos\left(\frac{1}{\sqrt{2^n}}\right)}{2\arcsin\left(\frac{1}{\sqrt{2^n}}\right)}$$

Since $\sin\alpha \approx \alpha$, this implies that $2\alpha \approx 2\sin\alpha = \frac{2}{\sqrt{2^n}}$. Then $2\alpha m \approx \frac{\pi}{2}$, so $m \approx \frac{\pi}{4\alpha} = \frac{\pi}{4}\sqrt{N}$. The number of iterations is independent of $|x_0\rangle$; it depends only on n.

Example. Consider a database with four items, so $n = 2, N = 4$. Here, $\sin\alpha = \frac{1}{2}$, so $\alpha = \frac{\pi}{6}$. Q causes a rotation through $2\alpha = \frac{\pi}{3}$. The initial state is

$$|\psi_0\rangle = |{+}{+}\rangle = \frac{1}{2}(|00\rangle + |01\rangle + |10\rangle + |11\rangle)$$

For any $x_0 \in B_2$, we have $\cos\beta = \langle x_0|\psi_0\rangle = \frac{1}{2}$ so $\beta = \frac{\pi}{3}$. Therefore, we need precisely one iteration, which rotates $|\psi_0\rangle$ to $|x_0\rangle$ exactly. Performing a measurement in the computational basis, we obtain x_0 with certainty.

We now prove the geometric properties of Q. First, note that $Q = -H_n I_0 H_n I_{x_0} = -I_{|\psi_0\rangle} I_{|x_0\rangle}$. If $|\alpha\rangle, |v\rangle \in \mathcal{V}$ and $|v\rangle \in \mathcal{P}(x_0)$, we have

- $I_{|x_0\rangle}|v\rangle = |v\rangle - 2\langle x_0|v\rangle|x_0\rangle$;
- $I_{|\psi_0\rangle}|v\rangle = |v\rangle - 2\langle\psi_0|v\rangle|\psi_0\rangle$.

These operators are reflections about lines perpendicular to $|x_0\rangle$ and $|\psi_0\rangle$ respectively. Thus, $\mathcal{P}(x_0)$ is stable under the action of $I_{|x_0\rangle}$ and $I_{|\psi_0\rangle}$.

Let M_1, M_2 be lines in the Euclidean plane, intersecting at O. Let θ be the angle between M_1 and M_2. Then, reflection about M_1 then M_2 acts as an anticlockwise rotation by 2θ about O.

In our case, the angle between the lines perpendicular to $|x_0\rangle$ and $|\psi_0\rangle$ is β. Therefore, $I_{|\psi_0\rangle}I_{|x_0\rangle}$ is an anticlockwise rotation by an angle of 2β. For any real unit vector $v \in \mathbb{R}^2$, we have $-I_v = I_{v^\perp}$ where $v_\perp$ is a unit vector orthogonal to v. Hence, $-I_{|\psi_0\rangle}I_{|x_0\rangle} = I_{|\psi_0^\perp\rangle}I_{|x_0\rangle}$, which is an anticlockwise rotation by an angle of 2α, as $\alpha + \beta = \frac{\pi}{2}$. This proves property (i).

Now consider $|\xi\rangle \in \mathcal{P}(x_0)^\perp$ perpendicular to $|\psi_0\rangle$ and to $|x_0\rangle$. Clearly $I_{|x_0\rangle}|\xi\rangle = |\xi\rangle$ and $I_{|\psi_0\rangle}|\xi\rangle = |\xi\rangle$. So $Q|\xi\rangle = -|\xi\rangle$, giving property (ii).

Grover's algorithm achieves an unstructured search for a unique good item in approximately $\frac{\pi}{4}\sqrt{N}$ queries, and there is no algorithm that has smaller asymptotic query complexity. Any quantum algorithm that achieves this search in an unstructured database of size N must use $O(\sqrt{N})$ queries. Moreover, it can be shown that $\frac{\pi}{4}(1-\varepsilon)\sqrt{N}$ queries are insufficient for each ε, so Grover's algorithm is tight.

4.10. Grover's algorithm for multiple items

Consider the case where there are $r \geq 1$ good items, and r is known. Here, $f(x_i) = 1$ if $i = 1, \dots, r$, and $f(x) = 0$ otherwise, where $x_1, \dots, x_r$ are the binary labels for the good items. We want to find any of the good items. Then, define

$$I_G |x\rangle = I - 2\sum_{i=1}^{r} |x_i\rangle\langle x_i| = \begin{cases} + |x\rangle & x \notin \{x_1, \dots, x_r\} \\ - |x\rangle & x \in \{x_1, \dots, x_r\} \end{cases}$$

Note that I_G is not of the form $I_{|v\rangle}$ for a single vector $|v\rangle$. Now, define $Q_G = -H_n I_0 H_n I_G = -I_{|\psi_0\rangle} I_G$. Let $|\psi_G\rangle = \frac{1}{\sqrt{r}} \sum_{i=1}^{r} |x_i\rangle$ be an equal superposition of the good states, and $|\psi_B\rangle = \frac{1}{\sqrt{N-r}} \sum_{i=r+1}^{N} |x_i\rangle$ be an equal superposition of the bad states. Note that $\langle\psi_G|\psi_B\rangle = 0$. Begin with the equal superposition state.

$$|\psi_0\rangle = (H|0\rangle)^{\otimes N} = \frac{\sqrt{r}}{\sqrt{N}} |\psi_G\rangle + \frac{\sqrt{N-r}}{\sqrt{N}} |\psi_B\rangle$$

Consider the plane $\mathcal{P}_G$ spanned by $|\psi_G\rangle$ and $|\psi_0\rangle$, which contains $|\psi_B\rangle$. Let α be the angle between $|\psi_G\rangle$ and $|\psi_0^\perp\rangle$.

We show that in the plane $\mathcal{P}_G$, Q_G acts as a rotation through an angle 2α where $\sin\alpha = \langle\psi_0|\psi_G\rangle = \frac{\sqrt{r}}{\sqrt{N}}$. The states $|\psi_G\rangle, |\psi_B\rangle$ form an orthonormal basis for $\mathcal{P}_G$. We find $I_G(a|\psi_G\rangle + b|\psi_B\rangle) = -a|\psi_G\rangle + b|\psi_B\rangle$; indeed, restricting to the plane $\mathcal{P}_G$, the action of I_G is precisely the action of $I_{|\psi_G\rangle}$. Hence, as before, Q_G causes the desired rotation through 2α in this plane. The probability of finding a single good item is $|\langle\psi|\psi_G\rangle|^2$, as $|\psi\rangle = a|\psi_G\rangle + b|\psi_B\rangle$.

Suppose now that r is unknown. In this case, we start with $|\psi_0\rangle$ and repeatedly apply Q to rotate $|\psi_0\rangle$ to $|\psi_G\rangle$ as before. However, we do not know how many iterations of Q to apply, since this depends on r.

If $r \ll N$, we choose K uniformly at random in $\left(0, \frac{\pi}{4}\sqrt{N}\right)$, and apply K iterations of Q. We measure the final state $|\psi^K\rangle$ to obtain x, and check if $f(x) = 1$ or not. Note that each iteration causes a rotation of 2α where $\sin\alpha = \frac{\sqrt{r}}{\sqrt{N}}$ so $2\alpha \approx 2\frac{\sqrt{r}}{\sqrt{N}}$. Choosing K therefore implicitly chooses a random angle in the range $\left(0, \frac{\pi}{2}\sqrt{r}\right)$. Now, if the final rotated state $|\psi\rangle$ makes an angle within $\pm\frac{\pi}{4}$ with $|\psi_0\rangle$, the probability of locating a good item is $|\langle\psi|\psi_0\rangle|^2 \geq \cos^2\frac{\pi}{4} = \frac{1}{2}$. Since for every quadrant in the plane $\mathcal{P}_G$, half of the angles are within $\pm\frac{\pi}{4}$ from the y-axis, the randomised procedure using $O(\sqrt{N})$ queries will locate a good item with probability approximately $\frac{1}{4}$. The procedure can then be repeated to reduce the error probability to an acceptable level.

4.11. NP problems

A *verifier* V for a language L is a computation with two inputs w, c such that

(i) if $w \in L$, there exists a *certificate of membership* c such that $V(w, c)$ halts in an accepting state; and

(ii) if $w \notin L$, for any c, $V(w, c)$ halts in a rejecting state.

V is a *poly-time* verifier if for all inputs w, c, the algorithm V runs in polynomial time in n, where n is the size of the input w. A problem in the *non-deterministic polynomial-time* complexity class NP is easy to verify, but may be hard to solve. More precisely, a language L is in NP if it has a polynomial time verifier V.

Alternatively, consider a computer operating non-deterministically; at each binary choice, the computer duplicates itself and performs both branches in parallel. We require that all possible paths eventually halt with either an accepting or rejecting state. The running time of a given algorithm is the length of the longest path. The computation is defined to accept its input if at least one path accepts it, and rejects its input if all paths reject it. One can check that NP is precisely the class of languages that are decided by a non-deterministic computation with polynomial running time.

Let $f : B_n \to B$ be a Boolean formula. The *Boolean satisfiability problem* SAT seeks an assignment of the variables $x_1, \dots, x_n$ such that $f(x_1, \dots, x_n) = 1$. Any such assignment is called a *satisfying assignment*. This problem clearly lies in NP; if f is satisfiable, then c is any assignment for which $V(f, c) = 1$ where $V(f, c) = f(c)$. Brute-force methods have $O(2^n)$ runtime.

Searching for arbitrarily many good items in an unstructured database corresponds to SAT. Assuming that there are few satisfying assignments, we can run the randomised Grover's algorithm to give a quantum algorithm for solving SAT in $O(\sqrt{2^n})$ time with low probability of error. Any NP problem can be converted into an application of SAT; we say SAT is NP-*complete*. Grover's algorithm can hence be applied to any NP problem to provide a quadratic speedup.

4.12. Shor's algorithm

Suppose N is a positive integer and $n = \lceil \log N \rceil$ is the number of bits in a binary representation of N. We wish to factorise N. We will describe an algorithm which, given N and a fixed acceptable probability of error, outputs a factor $1 < k < N$, or outputs N if N is prime. This algorithm runs in polynomial time in n; there is no classical algorithm with this property.

We first use results from number theory to convert the problem into a periodicity determination problem. Then, we apply the quantum period-finding algorithm using the quantum Fourier transform.

Choose an integer $1 < a < N$ uniformly at random, and compute $b = \gcd(a, N)$. If $b > 1$, then $b \mid N$ so is a factor; in this case we simply output b. If $b = 1$, then a, N are coprime.

Theorem (Euler's theorem)**.** Let a, N be coprime. Then there exists $1 < r < N$ such that $a^r \equiv 1 \bmod N$. A minimal such r is called the *order* of a modulo N.

Consider the *modular exponentiation function* $f : \mathbb{Z} \to {}^{\mathbb{Z}}\!/_{n\mathbb{Z}}$ such that $f(k) = a^k \bmod N$. This function satisfies $f(k_1 + k_2) = f(k_1)f(k_2)$. f is periodic with period r, and is injective within each period.

Suppose that we can find r, and suppose r is even. Then $a^r - 1 \equiv (a^{\frac{r}{2}} + 1)(a^{\frac{r}{2}} - 1) \equiv 0 \bmod N$. Note that $N \nmid (a^{\frac{r}{2}} - 1)$ since r was minimal such that $a^r \equiv 1 \bmod N$. If $N \nmid (a^{\frac{r}{2}} + 1)$, then N must have some prime factors in $(a^{\frac{r}{2}} + 1)$ and some in $(a^{\frac{r}{2}} - 1)$. We can use Euclid's algorithm to compute $\gcd(a^{\frac{r}{2}} + 1, N)$ and $\gcd(a^{\frac{r}{2}} - 1, N)$, which are factors of N. Thus, we find factors of N provided r is even and $a^{\frac{r}{2}} + 1 \not\equiv 0 \bmod N$.

Consider $N = 15, a = 7$. Then $f(k) = 7^k \bmod 15$ takes values $1, 7, 4, 13$, so has period $r = 4$. This is even, so we can write $a^r - 1 = (a^{\frac{r}{2}} + 1)(a^{\frac{r}{2}} - 1) = 50 \cdot 48$. $N = 15$ does not divide 50, so $\gcd(50, N) = 5$ is a factor, and $\gcd(48, 15) = 3$ is a factor.

Theorem. Let N be odd and not a prime power. Then, choosing a uniformly at random such that $\gcd(a, N) = 1$, the probability that r is even and $(a^{\frac{r}{2}} + 1) \not\equiv 0 \bmod N$ is at least $\frac{1}{2}$.

This implies that if N is odd and not a prime power, we obtain a factor of N with probability at least $\frac{1}{2}$. We repeat this process until the probability of not finding a factor is acceptably low. If N is even, we simply output 2 as a factor.

Lemma. Let $N = c^\ell$ for some $c, \ell \in \mathbb{N}$. There is a classical polynomial-time algorithm that computes c.

Shor's algorithm can be summarised as follows.

(i) Test if N is even; if so, output 2 and halt.

(ii) Run the classical algorithm to test if N is of the form c^ℓ with $\ell > 1$; if so, output c and halt.

(iii) Choose $1 < a < N$ uniformly at random and compute $b = \gcd(a, N)$. If $b > 1$, output b and halt.

(iv) Find the period r of the modular exponentiation function $f(k) = a^k \bmod N$. If this fails, return to step (iii).

(v) If r is even and $(a^{\frac{r}{2}} + 1) \not\equiv 0 \bmod N$, compute $t = \gcd(a^{\frac{r}{2}} + 1, N)$; if $1 < t < N$, output t and halt. Otherwise, return to step (iii).

We now describe the method to compute the period of the modular exponentiation function. Note that $f : \mathbb{Z} \to \mathbb{Z}$, not $\mathbb{Z}_N \to \mathbb{Z}_M$; we therefore cannot directly use the algorithm discussed previously. We must first truncate the domain $\mathbb{Z}$ to some $\mathbb{Z}_M$. If r is unknown, f will not necessarily be periodic on $\mathbb{Z}_M$. However, if M is $O(N^2)$, the single incomplete period has a negligible effect on the periodicity determination. We will define $M = 2^m$ for some m and use QFT_M.

Consider a finite domain $D = \{0, \dots, 2^m - 1\}$, where m is the smallest integer such that $2^m > N^2$. Suppose $2^m = Br + b$ where $0 \le b < r$, so $B = \left\lfloor \frac{2^m}{r} \right\rfloor$. We start with the equal superposition state $|\psi_m\rangle = \frac{1}{\sqrt{2^m}} \sum_{x \in D} |x\rangle$. Consider the quantum oracle U_f corresponding to the modular exponentiation function f. Then

$$\begin{aligned}|\Psi\rangle &= U_f |\psi_m\rangle |0\rangle \\ &= \frac{1}{\sqrt{2^m}} \sum_{x \in D} |x\rangle |f(x)\rangle \\ &= \frac{1}{\sqrt{2^m}} \sum_{x_0=0}^{b-1} \sum_{j=0}^{B} |x_0 + jr\rangle |f(x_0)\rangle + \frac{1}{\sqrt{2^m}} \sum_{x_0=b}^{r} \sum_{j=0}^{B-1} |x_0 + jr\rangle |f(x_0)\rangle\end{aligned}$$

Measuring the second register, we obtain an outcome $y = f(x_0)$. In the case $x_0 < b$, $f(x_0) = f(x_0 + jr)$ for $j \in \{0, \dots, B\}$. If $x_0 \ge b$, $f(x_0) = f(x_0 + jr)$ for $j \in \{0, \dots, B - 1\}$.

If $y = f(x_0)$ for $x_0 < b$, the probability of measuring y is $\frac{B+1}{2^m}$. The post-measurement state of the first register is $|\text{per}\rangle = \frac{1}{\sqrt{B+1}} \sum_{j=0}^{B} |x_0 + jr\rangle$. In the case $x_0 \ge b$, we have $|\text{per}\rangle = \frac{1}{\sqrt{B}} \sum_{j=0}^{B-1} |x_0 + jr\rangle$. In both cases,

$$|\text{per}\rangle = \frac{1}{\sqrt{A}} \sum_{j=0}^{A-1} |x_0 + jr\rangle$$

where $A = B + 1$ if $y = f(x_0)$ with $x_0 < b$ and $A = B$ if $y = f(x_0)$ with $x_0 \ge b$. We act on $|\text{per}\rangle$ by QFT_{2^m} to obtain

$$\begin{aligned}QFT_{2^m} |\text{per}\rangle &= \frac{1}{\sqrt{A}} \frac{1}{\sqrt{2^n}} \sum_{j=0}^{A-1} \sum_{c=0}^{2^m-1} \omega^{(x_0+jr)c} |c\rangle \\ &= \frac{1}{\sqrt{A}} \frac{1}{\sqrt{2^n}} \sum_{c=0}^{2^m-1} \omega^{x_0 c} \underbrace{\left[\sum_{j=0}^{A-1} (\omega^{cr})^j \right]}_{S} |c\rangle\end{aligned}$$

where $\omega = 2^{\frac{2\pi i}{M}}$ where $M = 2^m$. S is a geometric series. If $\frac{M}{r} \notin \mathbb{Z}$, $\alpha^A \neq 1$. We claim that a measurement on $QFT_{2^m} |\text{per}\rangle$ yields an integer c which is close to a multiple of $\frac{M}{r}$ with high probability.

Consider $k\frac{2^m}{r}$ for $k = 0, \dots, r - 1$. Each of these multiples is within $\frac{1}{2}$ of a unique integer; indeed, $2^m = Br + b$ so $r < 2^m$, giving that $k\frac{2^m}{r}$ cannot be a half integer. Consider the values of c such that $\left| c - k\frac{2^m}{r} \right| < \frac{1}{2}$ for $k = 0, \dots, r - 1$.

Theorem. Suppose that $QFT_{2^m}\left|\text{per}\right\rangle = \sum_{c=0}^{2^m-1} g(c)\left|c\right\rangle$, and that we measure the state and receive an outcome c. Let c_k be the unique integer such that $\left|c_k - k\frac{2^m}{r}\right| < \frac{1}{2}$. Then $\mathbb{P}\left(c = c_k\right) > \frac{\gamma}{r}$ for a fixed constant γ (which can be shown to be $\frac{4}{\pi^2}$). Moreover, the probability that k, r are coprime is $\Omega\left(\frac{1}{\log\log r}\right)$ by the coprimality theorem.

Thus, with $O(\log\log N) > O(\log\log r)$ repetitions, we obtain a good c value with high probability. Suppose that we measure c such that $\left|c - k\frac{2^m}{r}\right| < \frac{1}{2}$, so $\left|\frac{c}{2^m} - \frac{k}{r}\right| < \frac{1}{2^{m+1}}$. Recall that $r < N$ and m is minimal such that $2^m > N^2$. Then $\left|\frac{c}{2^m} - \frac{k}{r}\right| < \frac{1}{2N^2}$. Note that $\frac{c}{2^m}$ is known.

We show that there is at most one fraction $\frac{k}{r}$ with denominator $r < N$ such that $\left|\frac{c}{2^m} - \frac{k}{r}\right| < \frac{1}{2N^2}$. Suppose $\frac{k'}{r'}, \frac{k''}{r''}$ both satisfy this requirement. Then

$$\left|\frac{k'}{r'} - \frac{k''}{r''}\right| = \frac{\left|k'r'' - k''r'\right|}{r'r''} \geq \frac{1}{r'r''} > \frac{1}{N^2}$$

But $\left|\frac{c}{2^m} - \frac{k'}{r'}\right|, \left|\frac{c}{2^m} - \frac{k'}{r'}\right| < \frac{1}{2N^2}$, contradicting the triangle inequality. This result is the reason for choosing m minimal such that $2^m > N^2$. Therefore, we have with high probability that $\frac{c}{2^m}$ is close to a unique fraction $\frac{k}{r}$.

Example. Let $N = 39$ and choose $a = 7$; note that 7 and 39 are coprime. Let r be the period of $f(k) = a^k \bmod 39$. Note that $2^{10} < N^2 < 2^{11}$, so set $m = 11$. Suppose that $QFT_{2^{11}}\left|\text{per}\right\rangle$ gives a measurement of c. Then $\left|c - k\frac{2^{11}}{r}\right| < \frac{1}{2}$ with probability $\frac{\gamma}{r}$.

Suppose $c = 853$. One can explicitly check all fractions of the form $\frac{a}{b}$ to find one that satisfies $\left|\frac{a}{b} - \frac{853}{2048}\right| < \frac{1}{2^{12}}$. This is consistent with $\frac{a}{b} = \frac{5}{12}, \frac{10}{24}$; as we are constrained by coprimality we must choose $r = 12$. One can check that $7^{12} \equiv 1 \bmod 39$, hence $r = 12$. Note that $O(N^2) = O(\exp(n))$ computations are needed for this calculation; there is a more efficient way to compute a, b using continued fractions.

A rational number $\frac{s}{t}$ can be written in the form of a continued fraction

$$\frac{s}{t} = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\dots + \cfrac{1}{a_\ell}}}} = [a_1, \dots, a_\ell]$$

where $a_1, \dots, a_\ell$ are positive integers. We can write $\frac{s}{t} = \frac{1}{\frac{t}{s}} = \frac{1}{a_1 + \frac{s_1}{t_1}}$, and so on. For example, if $\frac{s}{t} = \frac{13}{35}$, we can find $a_1 = 2, a_2 = 2, a_3 = 1, a_4 = 1, a_5 = 2$ and $\ell = 5$. Since the sequence t_k is decreasing, the expansion will always terminate. For each $k = 1, \dots, \ell$, we can truncate the computation at level k. This gives the sequence of rational numbers

$$\frac{p_1}{q_1} = [a_1], \frac{p_2}{q_2} = [a_1, a_2], \dots, \frac{p_\ell}{q_\ell} = [a_1, \dots, a_\ell] = \frac{s}{t}$$

$\frac{p_k}{q_k}$ is the kth *convergent* of the continued fraction $\frac{s}{t}$.

Lemma. Let $a_1, \dots, a_\ell$ be positive reals, and let $p_0 = 0, q_0 = 1, p_1 = 1, q_1 = a_1$. Then,

(i) $[a_1, \dots, a_k] = \frac{p_k}{q_k}$ where $p_k = a_k p_{k-1} + p_{k-2}$ and $q_k = a_k q_{k-1} + q_{k-2}$;

(ii) if the a_k are integers, then so are the p_k and q_k, with $q_k p_{k-1} - p_k q_{k-1} = (-1)^k$ for $k \geq 1$, and moreover $\gcd(p_k, q_k) = 1$.

Theorem. Consider a continued fraction $\frac{s}{t} = [a_1, \dots, a_\ell]$, and let $\frac{p_k}{q_k}$ be the kth convergent. If s and t are given by m-bit integers, then the length ℓ of the continued fraction is $O(m)$, and the continued fraction and its convergents can be computed in $O(m^3)$ time.

Proof sketch. We have $a_k \geq 1$ and $p_k, q_k \geq 1$. Part (i) of the above lemma implies that (p_k) and (q_k) are increasing sequences. If k is even, $p_k \geq 2p_{k-2}$ and $q_k \geq 2q_{k-2}$ hence $p_k, q_k \geq 2^{\frac{k}{2}}$. Thus, in general, $p_k, q_k \geq 2^{\left\lfloor \frac{k}{2} \right\rfloor}$. We therefore need at most $\ell = O(m)$ iterations to obtain $\frac{s}{t}$ exactly, since q_k, p_k are coprime and each are at least $2^{\left\lfloor \frac{k}{2} \right\rfloor}$. The computation of each successive a_k value involves division of $O(m)$-bit integers and converting it into an integer and remainder term; these computations can be performed in $O(m^2)$ time. Hence, the entire computation requires only $O(m^3)$ time. □

Theorem. Let $x \in \mathbb{Q}$ with $0 < x < 1$. Let $\frac{p}{q} \in \mathbb{Q}$ such that $\left| x - \frac{p}{q} \right| < \frac{1}{2q^2}$. Then $\frac{p}{q}$ is a convergent of the continued fraction expansion of x.

In our situation, we have c such that

$$\left| \frac{c}{2^m} - \frac{k}{r} \right| < \frac{1}{2N^2}; \quad r < N$$

In particular, $\left| \frac{c}{2^m} - \frac{k}{r} \right| < \frac{1}{2r^2}$, and we have seen that there is at most one fraction $\frac{k}{r}$ such that this holds. Note that $0 < c < 2^m$, so $0 < \frac{c}{2^m} < 1$. Hence, $\frac{k}{r}$ is a convergent of $\frac{c}{2^m}$. Note that $2^m > N^2 > 2^{m-1}$, so $c, 2^m$ are $O(m)$-bit integers, and hence the sequence of convergents (and in particular $\frac{k}{r}$) can be computed in $O(m^3)$ time. We can then explicitly check for each convergent $\frac{k}{r}$ if $\left| \frac{c}{2^m} - \frac{k}{r} \right| < \frac{1}{2N^2}$ and $r < N$ hold.

Example. Consider again $N = 39$ and $2^m = 2^{11} = 2048$. Suppose $c = 853$. Then one can explicitly compute

$$\frac{c}{2^m} = \frac{853}{2048} = [2, 2, 2, 42, 4]$$

Its convergents are

$$\frac{1}{2}; \quad \frac{2}{5}; \quad \frac{5}{12}; \quad \frac{212}{509}; \quad \frac{853}{2048}$$

Only $\frac{5}{12}$ satisfies $\left| \frac{c}{2^m} - \frac{k}{r} \right| < \frac{1}{2N^2}$ and $r < N$. So $r = 12$ is the period.

A classical factoring algorithm takes $O\left(\exp\left(n^{\frac{1}{3}}\right)\right)$ time; we analyse the time complexity of Shor's algorithm. Consider the case when N is odd and not a prime power, and let $n = \log N$. The modular exponentiation function requires $O(m) = O(n)$ multiplications, each of which take $O(m^2) = O(n^2)$ time, so this algorithm takes $O(n^3)$ time. The construction of the equal superposition state requires $m = O(n)$ Hadamard gates, and applying the quantum oracle gives the state $\frac{1}{2^m} \sum_{x \in B_m} |x\rangle |f(x)\rangle$ in $O(n^3)$ steps. We measure the second register which contains $O(n)$ qubits, hence requiring $O(n)$ single-qubit measurements. The first register is then in state $|\text{per}\rangle$. We then apply the quantum Fourier transform QFT_{2^m}, which can be implemented in $O(m^2) = O(n^2)$ steps. We then measure the first register to obtain c, requiring $O(n)$ single-qubit measurements. Then, we find r from c using the continued fraction algorithm, requiring $O(n^3)$ steps. A good c value is obtained with probability $1 - \varepsilon$ with $O(\log \log N) = O(\log n)$ repetitions. Then, $t = \gcd(a^{\frac{r}{2}} + 1, N)$ is computed using Euclid's algorithm, taking $O(n^3)$ steps. If r is odd or is even but $t = 1$, then we return to the start, and the case where r is even and $t \neq 1$ occurs with probability at least $1 - \varepsilon$ if we perform $\log \frac{1}{\varepsilon}$ repetitions.

VIII. Number Fields

Lectured in Lent 2023 by PROF. I. GROJNOWSKI

A number field is a field extension of $\mathbb{Q}$, generated by finitely many algebraic numbers. Common number fields include the rationals $\mathbb{Q}$, the Gaussian rationals $\mathbb{Q}(i)$, and more general quadratic fields $\mathbb{Q}(\sqrt{d})$. The field of rationals contain the ring of integers $\mathbb{Z}$, and each number field similarly contains its own ring of algebraic integers. For example, the ring of algebraic integers in $\mathbb{Q}(i)$ is $\mathbb{Z}[i]$.

While the field structure of a number field is typically relatively simple, the ring structure of the algebraic integers can offer more insight into the field in question. In general, the ring of algebraic integers is not a unique factorisation domain, but some fields have 'better' or 'worse' factorisation properties than others. We quantify the degree to which unique factorisation fails by assigning a group to each number field, called the class group. If the class group is large, there are many ways in which unique factorisation could fail. A remarkable fact proven in this course is that the class group is finite.

We also study the units in the ring of algebraic integers. The roots of unity in a number field are units, but there may be other units not of this form. Dirichlet's unit theorem describes a geometric interpretation of the set of units. In particular, modulo the roots of unity, they form a lattice isomorphic to $\mathbb{Z}^k$ for some $k \in \{0, 1, \dots\}$. We can use this theorem to find all of the integer solutions of problems like Pell's equation, $x^2 - dy^2 = \pm 1$.

Contents

1. Number fields

1.1. Algebraic integers

Recall that if K and L are fields and $\dim_K L < \infty$, we write $[L : K]$ for this dimension and say that L/K is a finite extension. If L/K is a finite extension, every element $x \in L$ is algebraic over K.

Definition. A *number field* is a finite extension of $\mathbb{Q}$.

Definition. Let L be a number field. $\alpha \in L$ is an *algebraic integer* if there exists $f \in \mathbb{Z}[x]$ monic such that $f(\alpha) = 0$. We write $\mathcal{O}_L = \{\alpha \in L \mid \alpha \text{ is an algebraic integer}\}$ for the set of *integers of L*.

$$\begin{tikzcd}
\mathbb{Z} \arrow[r] \arrow[d] & \mathbb{Q} \arrow[d] \\
\mathcal{O}_L \arrow[r] & L
\end{tikzcd}$$

Lemma. $\mathcal{O}_{\mathbb{Q}} = \mathbb{Z}$.

Proof. Clearly if α is an integer, then $f(x) = x - \alpha$ is a monic polynomial such that $f(\alpha) = 0$. Conversely, if α is a rational number, we can let $\alpha = \frac{r}{s}$ where r and s are coprime. Let $f(x) = x^n + a_{n-1}x^{n-1} + \dots + a_0 \in \mathbb{Z}[x]$ such that $f(\alpha) = 0$. Clearing denominators, $r^n + a_{n-1}r^{n-1}s + \dots + a_0 s^n = 0$. Hence $s \mid r^n$. If $s \neq 1$, let $p \mid s$ be a prime, then $p \mid r$, so r and s were not coprime. □

We will soon show that $\mathcal{O}_L$ is a ring. In other words, $\alpha, \beta \in \mathcal{O}_L$ implies $\alpha \pm \beta, \alpha\beta \in \mathcal{O}_L$.

Note that $\alpha \in \mathcal{O}_L$ does not in general imply $\frac{1}{\alpha} \in \mathcal{O}_L$. Recall from Galois Theory that if $\alpha, \beta \in L$, and α, β are algebraic over K, then so is $\alpha \pm \beta, \alpha\beta$. The proof from Galois Theory will not work in this case, since that proof does not provide for monic polynomials.

Definition. Let $R \subseteq S$ be commutative rings with a 1.

(i) $\alpha \in S$ is *integral over R* if there exists a monic polynomial $f \in R[x]$ such that $f(\alpha) = 0$.

(ii) S is *integral over R* if all $\alpha \in S$ are integral over R.

(iii) S is *finitely generated over R* if there exist elements $\alpha_1, \dots, \alpha_n \in S$ such that any element of S can be written as an R-linear combination of the α_i. Equivalently, the map $R^n \to S$ given by $(r_1, \dots, r_n) \mapsto \sum_{i=1}^n r_i \alpha_i$ is surjective.

Example. Let $\mathbb{Q} \subseteq L$ be a number field. Then $\alpha \in L$ is an algebraic integer if and only if α is integral over $\mathbb{Z}$. $\mathcal{O}_L$ is integral over $\mathbb{Z}$ (once we have proven it is a ring).

If $\alpha_1, \dots, \alpha_r \in S$, we write $R[\alpha_1, \dots, \alpha_r]$ for the subring of S generated by R and the α_i. This is equivalently the image of the polynomial ring $R[x_1, \dots, x_r] \to S$ mapping x_i to α_i.

Proposition. Let $S = R[s]$, where s is integral over R. Then S is finitely generated over R. Further, if $S = R[s_1, \dots, s_n]$ with each s_i integral over R, then S is finitely generated over R.

Proof. S is spanned by $1, s, s^2, \dots$ over R. By assumption, there exists $a_0, \dots, a_{n-1} \in R$ such that $s^n = \sum_{i=0}^{n-1} a_i s^i$. So the R-module spanned by $1, \dots, s^{n-1}$ is stable under multiplication by s, so contains $s^n, s^{n+1}, \dots$ and hence is all of S.

Let $S_i = R[s_1, \dots, s_{i-1}]$. Then $S_{i+1} = S_i[s_{i+1}]$, and s_{i+1} is integral over R, hence is integral over S_i. So S_{i+1} is finitely generated over S_i. Note that if $A \subseteq B \subseteq C$ where B is finitely generated over A and C is finitely generated over B, then C is finitely generated over A. Indeed, if b_i generate B over A and c_j generate C over B, the $b_i c_j$ generate C over A. □

Theorem. If S is finitely generated over R, S is integral over R.

Proof. Let $\alpha_1, \dots, \alpha_n$ generate S as an R-module. Without loss of generality, we can assume $\alpha_1 = 1$. Let $s \in S$, and consider the function $m_s : S \to S$ given by $m_s(x) = sx$. Then, $m_s(\alpha_i) = s\alpha_i = \sum b_{ij}\alpha_j$ for some choice of b_{ij}. Let $B = (b_{ij})$. By definition, $(sI - B)(\alpha_1, \dots, \alpha_n)^\intercal = 0$.

Recall that for any matrix X, the adjugate has the property that $\operatorname{adj}(X)X = \det X \cdot I$. Hence, $\det(sI - B)(\alpha_1, \dots, \alpha_n)^\intercal = 0$. In particular, $\det(sI - B)\alpha_1 = \det(sI - B) = 0$. Let $f(t) = \det(tI - B)$, which is a monic polynomial in R. As $f(s) = 0$, s is integral over R. □

Note the similarity to a proof of the Cayley–Hamilton theorem. Note further that this proof is constructive.

Corollary. Let $\mathbb{Q} \subseteq L$ be a number field. Then $\mathcal{O}_L$ is a ring.

Proof. If $\alpha, \beta \in \mathcal{O}_L$, then $\mathbb{Z}[\alpha, \beta]$ is finitely generated over $\mathbb{Z}$. So this ring is integral. □

Corollary. Let $A \subseteq B \subseteq C$ be ring extensions, where B/A is integral and C/B is integral. Then C/A is integral.

Proof. If $c \in C$, let $f(x) = \sum_{i=0}^{n} b_i x^i$ be the monic polynomial in $B[x]$ it satisfies, and set $B_0 = A[b_0, \dots, b_{n-1}]$, $C_0 = B[c]$. Then B_0 is finitely generated over A as $b_0, \dots, b_{n-1}$ are integral over A, and C_0 is finitely generated over B_0 as c is integral over B_0. C_0 is therefore finitely generated over A. Then the theorem implies that c is integral over A. □

Remark. C could have had infinitely many generators, for instance,

$$C = \{\alpha \in \mathbb{C} \mid \alpha \text{ is an algebraic integer}\}$$

This possibility is why we passed to C_0. This kind of proof is common in commutative algebra, applying a powerful theorem such as the Cayley–Hamilton theorem carefully to find its consequences.

Example. $\mathcal{O}_{\mathbb{Q}[i]} = \mathbb{Z}[i]$.

1.2. Minimal polynomials

Let $K \subseteq L$ be fields. Recall that the minimal polynomial of $\alpha \in L$ is the monic polynomial $p_\alpha(x) \in K[x]$ of minimum degree such that $p_\alpha(\alpha) = 0$.

Lemma. Let $f(x) \in K[x]$ satisfy $f(\alpha) = 0$. Then $p_\alpha \mid f$.

Proof. By Euclid, $f = p_\alpha h + r$ where $r \in K[x]$ has degree less than that of p. Then $0 = f(\alpha) = p_\alpha(\alpha)h(\alpha) + r(\alpha)$. If $r \neq 0$, this contradicts minimality of $\deg p_\alpha$. □

The converse is obvious, so the lemma implies the uniqueness of p_α.

Proposition. Let L be a number field and $\alpha \in L$. Then $\alpha \in \mathcal{O}_L$ if and only if $p_\alpha(x) \in \mathbb{Q}[x]$ is in $\mathbb{Z}[x]$.

Proof. If p_α has integer coefficients, this holds by definition. Conversely, suppose $\alpha \in \mathcal{O}_L$, where p_α is the minimal polynomial. Let $M \supseteq L$ be a splitting field for p_α, i.e. a field in which p_α splits into linear factors. Let $h(x)$ be a monic polynomial which α satisfies. By the lemma, $p_\alpha \mid h$, so each root α_i of p_α in M is an algebraic integer. By the previous theorem, sums and products of algebraic integers are algebraic. So the coefficients of p_α are algebraic integers. But $p_\alpha \in \mathbb{Q}[x]$, so the coefficients are in $\mathbb{Z}$. □

Remark. One can also show this from the previous result and Gauss' lemma.

Lemma. The field of fractions of $\mathcal{O}_L$ is L. In fact, if $\alpha \in L$, there exists $n \in \mathbb{Z}, n \neq 0$ such that $n\alpha \in \mathcal{O}_L$.

Proof. Let $\alpha \in L$, and g be the minimal polynomial of α. Then g is monic, and there exists an integer $n \in \mathbb{Z}, n \neq 0$ such that $ng \in \mathbb{Z}[x]$. So $h(x) = n^{\deg g} g\left(\frac{x}{n}\right)$ is an integer polynomial which is monic, and this is the minimal polynomial of $n\alpha$, so $n\alpha \in \mathcal{O}_L$. □

1.3. Integral basis

If L/K is a field extension, and $\alpha \in L$, we write $m_\alpha : L \to L$ for the map given by multiplication by α. We define the *norm* of α to be the determinant of m_α, and the *trace* of α to be the trace of m_α. Recall that if $p_\alpha(x)$ is the minimal polynomial of α, then the characteristic polynomial of m_α is $\det(xI - m_\alpha) = p_\alpha^{[L/K(\alpha)]}$. Further, if $p_\alpha(t)$ splits as $(t-\alpha_1)\cdots(t-\alpha_r)$ in some field $L' \supseteq K(\alpha)$, then $N_{K(\alpha)/K}(\alpha) = \prod \alpha_i$ and $\mathrm{Tr}_{K(\alpha)/K}(\alpha) = \sum \alpha_i$, and $N_{L/K}(\alpha) = (\prod \alpha_i)^{[L:K(\alpha)]}, \mathrm{Tr}_{L/K}(\alpha) = [L : K(\alpha)] \sum \alpha_i$.

If L is a number field, then α is an algebraic integer if and only if the minimal polynomial is has integer coefficients, which is the case if and only if the characteristic polynomial of m_α has integer coefficients. In particular, in this case, $N_{L/\mathbb{Q}}(\alpha) \in \mathbb{Z}$ and $\mathrm{Tr}_{L/\mathbb{Q}}(\alpha) \in \mathbb{Z}$. If the degree of L over $\mathbb{Q}$ is 2, the norm and trace are integers if and only if α is algebraic, since these values determine the characteristic polynomial.

Example. Let $L = K(\sqrt{d})$ where $d \in K$ is not a square. This has basis $1, \sqrt{d}$. If $\alpha = x + y\sqrt{d}$, the matrix m_α is

$$\begin{pmatrix} x & dy \\ y & x \end{pmatrix}$$

Then, $\mathrm{Tr}_{L/K}(x + y\sqrt{d}) = 2x = (x + y\sqrt{d}) + (x - y\sqrt{d})$, and $N_{L/K}(x + y\sqrt{d}) = x^2 - dy^2 = (x + y\sqrt{d})(x - y\sqrt{d})$.

Lemma. Let $L = \mathbb{Q}(\sqrt{d})$, $d \in \mathbb{Z}$ a nonzero square-free integer. Such a field is called a *quadratic field*. Then, $\mathcal{O}_L = \mathbb{Z}[\sqrt{d}]$ if $d \equiv 2, 3 \bmod 4$, and $\mathcal{O}_L = \mathbb{Z}\left[\frac{1}{2}(1 + \sqrt{d})\right]$ if $d \equiv 1 \bmod 4$.

Proof. $x + y\sqrt{d} \in \mathcal{O}_L$ if and only if $2x, x^2 - dy^2 \in \mathbb{Z}$. This implies that $4dy^2 \in \mathbb{Z}$. If $y = \frac{r}{s}$ with $\gcd(r, s) = 1$, then $s^2 \mid 4d$. But d was square-free, so $s^2 \mid 4$ so $s = \pm 1, \pm 2$. As $2x \in \mathbb{Z}$, we can write $x = \frac{u}{2}$ and $y = \frac{v}{2}$, for $u, v \in \mathbb{Z}$. Therefore, $u^2 - dv^2 \in 4\mathbb{Z}$, so $u^2 \equiv dv^2 \bmod 4$. Note that u^2 must be 0 or 1 mod 4.

So if d is not congruent to 1 mod 4, $u^2 \equiv dv^2$ has a solution, so u^2, v^2 are both zero mod 4, so u, v are even. In this case, $x, y \in \mathbb{Z}$, so any $\alpha \in \mathcal{O}_L$ is a $\mathbb{Z}$-combination of $1, \sqrt{d}$.

On the other hand, if $d \equiv 1$, then u, v have the same parity mod 2, so we can write any such values as a $\mathbb{Z}$-combination of $1, \frac{1}{2}(1 + \sqrt{d})$. □

Example. If $d = -1$, $\mathcal{O}_{\mathbb{Q}[i]} = \mathbb{Z}[i]$. Note that the minimal polynomial of $\frac{1}{2}(1 + \sqrt{d})$ is $t^2 - t + \frac{1}{4}(1 - d)$, which has integer coefficients as $d \equiv 1$.

Definition. Let L be a number field. Then, a basis $\alpha_1, \dots, \alpha_n$ of L as a $\mathbb{Q}$-vector space is called an *integral basis* if $\mathcal{O}_L = \left\{\sum_{i=1}^n m_i\alpha_i \mid m_i \in \mathbb{Z}\right\} = \bigoplus_{i=1}^n \mathbb{Z}\alpha_i$.

Example. $\mathbb{Q}(\sqrt{d})$ has integer basis $1, \frac{1}{2}(1 + \sqrt{d})$ or $1, \sqrt{d}$, depending on the value of d mod 4.

Integral bases are not unique. Given two such bases, there exists a matrix $g \in GL_n(\mathbb{Z})$ which transforms one into the other. We now aim to show that there exists an integral basis for every number field.

Recall that if L/K is a finite separable extension, then there exists $\alpha \in L$ such that $L = K(\alpha)$; this is the primitive element theorem. Note that all extensions in characteristic 0 are separable.

Example. $\mathbb{Q}(\sqrt{2}, \sqrt{3}) = \mathbb{Q}(\sqrt{2} + \sqrt{3})$.

This implies that if $L/\mathbb{Q}$ is a number field, then there exists $\alpha \in L$ such that $L = \mathbb{Q}(\alpha)$, isomorphic to ${}^{\mathbb{Q}[x]}\!/\!_{(p_\alpha(x))}$ where p_α is the minimal polynomial for x. L is a field, so $P_\alpha \lhd \mathbb{Q}[x]$ is a maximal ideal in the principal ideal domain $\mathbb{Q}[x]$, and p_α is irreducible. Let $\deg p_\alpha = [L : \mathbb{Q}] = n$. Then L has basis $1, \alpha, \dots, \alpha^{n-1}$ as a $\mathbb{Q}$-vector space.

Lemma. n is the number of field embeddings of L into $\mathbb{C}$.

Proof. $p_\alpha \in \mathbb{Q}[x]$ is irreducible, so $\gcd(p_\alpha, p'_\alpha) = 1$. So $p_\alpha(x) = (x - \alpha_1) \dots (x - \alpha_n)$ has n distinct roots in $\mathbb{C}$. A field homomorphism ${}^{\mathbb{Q}[x]}\!/\!_{(p_\alpha(x))} \to \mathbb{C}$ is automatically $\mathbb{Q}$-linear, so must map x to a root α_i of $p_\alpha(x)$ in $\mathbb{C}$. Conversely, there exists such a map for each α_i, and they are distinct. □

This allows us to define a new invariant which refines $n = [L : \mathbb{Q}]$.

Definition. Let r be the number of real roots of $p_\alpha(x)$, and let s be the number of complex conjugate pairs of roots of $p_\alpha(x)$. Also, r is the number of field embeddings of L into $\mathbb{R}$, so is independent of the choice of α. s is therefore also an invariant, as $r + 2s = n$.

Lemma. Let $L/\mathbb{Q}$ be a number field. Let $\sigma_1, \dots, \sigma_n : L \to \mathbb{C}$ be the different field embeddings, so $n = [L : \mathbb{Q}]$. If $\beta \in L$, then $\mathrm{Tr}_{L/\mathbb{Q}}(\beta) = \sum \sigma_i(\beta)$ and $N_{L/\mathbb{Q}}(\beta) = \prod \sigma_i(\beta)$. We call the $\sigma_i(\beta)$ the *conjugates* of β in $\mathbb{C}$.

Example. If $L = \mathbb{Q}(\sqrt{d})$ where d is square-free, then $a + b\sqrt{d}$ and $a - b\sqrt{d}$ are conjugates.

Proposition. Let L/K be a finite separable extension. Then, the K-bilinear form $L \times L \to K$ given by $(x, y) \mapsto \mathrm{Tr}_{L/K}(xy)$, known as the *trace form*, is a nondegenerate inner product. Equivalently, if $\alpha_1, \dots, \alpha_n$ is a basis of L/K, the Gram matrix has nonzero determinant, that is, $\Delta(\alpha_1, \dots, \alpha_n) = \det \mathrm{Tr}_{L/K}(\alpha_i\alpha_j) \neq 0$. Conversely, if L/K is inseparable, the trace form is the zero map.

Proof. Let $\sigma_1, \dots, \sigma_n : L \to \overline{K}$ be the n distinct K-linear field embeddings of L into an algebraic closure $\overline{K}$, which exists by separability. Let S be the matrix $(\sigma_i(\alpha_j))$. Observe that $S^\intercal S$ is the matrix with (i, j) term

$$\sum_{k=1}^n \sigma_k(\alpha_i)\sigma_k(\alpha_j) = \sum_{k=1}^n \sigma_k(\alpha_i\alpha_j) = \mathrm{Tr}_{L/K}(\alpha_i\alpha_j)$$

So $\Delta(\alpha_1, \dots, \alpha_n) = \det S \det S^\intercal = (\det S)^2$. By the primitive element theorem, there exists $\theta \in L$ such that $L = K(\theta)$. Therefore, $1, \theta, \dots, \theta^{n-1}$ forms a basis of L/K. Then

$$S = \begin{pmatrix} 1 & \sigma_1(\theta) & \cdots & \sigma_1(\theta^{n-1}) \\ \vdots & \vdots & & \vdots \\ 1 & \sigma_n(\theta) & \cdots & \sigma_n(\theta^{n-1}) \end{pmatrix}$$

This is a Vandermonde matrix, which gives

$$(\det S)^2 = \prod_{i<j} \big(\sigma_i(\theta) - \sigma_j(\theta)\big)^2 = \Delta(1, \theta, \dots, \theta^{n-1})$$

This is nonzero; indeed, if $\sigma_i(\theta) = \sigma_j(\theta)$, then $\sigma_i(\theta^a) = \sigma_j(\theta^a)$ for all a, so $\sigma_i = \sigma_j$, but they are distinct.

Moreover, if $\alpha_1, \dots, \alpha_n$ is any basis of L/K, and $\alpha_1', \dots, \alpha_n'$ is another basis of L/K, then

$$\Delta(\alpha_1', \dots, \alpha_n') = (\det A)^2 \Delta(\alpha_1, \dots, \alpha_n)$$

where $\alpha_i' = \sum a_{ij}\alpha_j$ and $A = (a_{ij})$. Hence, $\Delta(\alpha_1, \dots, \alpha_n) \neq 0$ for any basis. □

Remark. $L = K(\theta)$ and $p_\theta(t) = \prod(t - \sigma_i(\theta))$. The Galois theory notion of the discriminant of p_θ, which is $\prod_{i<j}(\sigma_i(\theta) - \sigma_j(\theta))^2$, is exactly the determinant of the Gram matrix $\Delta(1, \theta, \dots, \theta^{n-1})$, also often called a discriminant.

Remark. Let L be a number field. If $\alpha, \beta \in \mathcal{O}_L$, $\operatorname{Tr}_{L/\mathbb{Q}}(\alpha\beta) \in \mathbb{Z}$. Therefore, the inner product is a function $\mathcal{O}_L \times \mathcal{O}_L \to \mathbb{Z}$. If $\alpha_1, \dots, \alpha_n \in L$ form a basis of L over $\mathbb{Q}$, and $\alpha_1, \dots, \alpha_n$ are algebraic integers, then $\Delta(\alpha_1, \dots, \alpha_n)$ is a nonzero integer.

Theorem. Let $L/\mathbb{Q}$ be a number field. Then there exists an integral basis for $\mathcal{O}_L$: there exist $\alpha_1, \dots, \alpha_n \in \mathcal{O}_L$ such that $\mathcal{O}_L = \bigoplus \mathbb{Z}\alpha_i \simeq \mathbb{Z}^n$ and $L = \bigoplus \mathbb{Q}\alpha_i \simeq \mathbb{Q}^n$.

Proof. Let $\alpha_1, \dots, \alpha_n$ be any basis for L as a $\mathbb{Q}$-vector space. We have shown that there exists $n_i \in \mathbb{Z}$ such that $n_i\alpha_i \in \mathcal{O}_L$. Therefore, we can assume $\alpha_1, \dots, \alpha_n \in \mathcal{O}_L$ without loss of generality. Here, $\Delta(\alpha_1, \dots, \alpha_n)$ is a nonzero integer.

Choose $\alpha_1, \dots, \alpha_n$ such that $\Delta(\alpha_1, \dots, \alpha_n)$ has minimum absolute value. Suppose the result is false, so let $x \in \mathcal{O}_L$ and $x = \sum \lambda_i\alpha_i$ where $\lambda_i \in \mathbb{Q}$, and suppose that some λ_i is not an integer. Without loss of generality let $\lambda_1 \notin \mathbb{Z}$. Write $\lambda_1 = n_1 + \varepsilon_1$, and $0 < \varepsilon_1 < 1$. Now, let

$$\alpha_1' = x - n_1\alpha_1 = \varepsilon_1\alpha_1 + \lambda_2\alpha_2 + \dots + \lambda_n\alpha_n$$

Note $\alpha_1' \in \mathcal{O}_L$. Then $\alpha_1', \alpha_2, \dots, \alpha_n$ is a basis of L containing only the elements of $\mathcal{O}_L$. But $\Delta(\alpha_1', \alpha_2, \dots, \alpha_n) = \varepsilon_1^2\Delta(\alpha_1, \dots, \alpha_n)$ contradicting the minimality assumption. □

Remark. If $\alpha_1', \dots, \alpha_n'$ are any other integral basis of $\mathcal{O}_L$, then there exists $g \in GL_n(\mathbb{Z})$ such that $g(\alpha_i') = \alpha_i$. But $\det g \in GL_1(\mathbb{Z}) = \{\pm 1\}$, so $(\det g)^2 = 1$, giving $\Delta(\alpha_1', \dots, \alpha_n') = \Delta(\alpha_1, \dots, \alpha_n)$, so this is an invariant.

Definition. The *discriminant* of a number field $L/\mathbb{Q}$ is the invariant $D_L = \Delta(\alpha_1, \dots, \alpha_n)$.

Example. Let $L = \mathbb{Q}(\sqrt{d})$ where d is square-free. Then, $d \equiv 2, 3 \bmod 4$, then $1, \sqrt{d}$ is an integral basis. If $d \equiv 1 \bmod 4$, then $1, \frac{1}{2}\left(1 + \sqrt{d}\right)$ is an integral basis. Then,

$$D_L = \left[\det\begin{pmatrix} 1 & \sqrt{d} \\ 1 & -\sqrt{d} \end{pmatrix}\right]^2 = 4d; \quad D_L = \left[\det\begin{pmatrix} 1 & \frac{1}{2}\left(1 + \sqrt{d}\right) \\ 1 & \frac{1}{2}\left(1 - \sqrt{d}\right) \end{pmatrix}\right]^2 = d$$

So the discriminant is either $4d$ or d.

Remark. Results on quadratic fields are often phrased more uniformly if written in terms of D_L. Note also that $L = \mathbb{Q}(\sqrt{D_L})$. An integral basis is $1, \frac{\sqrt{D_L}+D_L}{2}$ regardless of the value of d.

2. Ideals

2.1. Ideals in the ring of integers

Lemma. Let $x \in \mathcal{O}_L$, where L is a number field. Then x is a unit in $\mathcal{O}_L$ if and only if $N_{L/\mathbb{Q}}(x) = \pm 1$. We write $\mathcal{O}_L^\star$ for the set of units of $\mathcal{O}_L$.

Proof. If x is a unit, then as the norm is multiplicative, $N(xx^{-1}) = 1$ so $N(x)N(x^{-1}) = 1$. So $N(x) = \pm 1$. Conversely, let $\sigma_1, \dots, \sigma_n : L \to \mathbb{C}$ be the distinct field embeddings. Let $L \subseteq \mathbb{C}$ be the containment given by σ_1. If $x \in \mathcal{O}_L$, then $N(x) = x\sigma_2(x) \dots \sigma_n(x)$. So if $N(x) = \pm 1$, we have $\frac{1}{x} = \pm \prod_{i=2}^n \sigma_i(x)$. This is a product of algebraic integers, hence an algebraic integer. So $x^{-1} \in \mathcal{O}_L$. □

Recall that if $x \in \mathcal{O}_L$, it is irreducible if it does not factorise as ab where $a, b \in \mathcal{O}_L$ not units. If $x = uy$ where u is a unit, we say x and y are associate. Many number fields have rings of algebraic integers which are not unique factorisation domains.

Example. Let $L = \mathbb{Q}\big(\sqrt{-5}\big)$. Here, $\mathcal{O}_L = \mathbb{Z}\big[\sqrt{-5}\big]$. Note that $3 \cdot 7 = \big(1 + 2\sqrt{-5}\big)\big(1 - 2\sqrt{-5}\big)$, and $N(3) = 9, N(7) = 49, N\big(1 \pm \sqrt{-5}\big) = 21$. These are not associates. We claim that $3, 7, 1 \pm 2\sqrt{-5}$ are irreducible, so $\mathcal{O}_L$ is not a unique factorisation domain. If this were not the case, $3 = \alpha\overline{\alpha}$, where $\alpha = x + y\sqrt{-5}$, but $N(3) = 9 = N(\alpha)N(\overline{\alpha}) = N(\alpha)^2$ so $N(\alpha) = x^2 + 5y^2 = \pm 3$, but there are no integer solutions to this equation. All of the other factors are similarly irreducible.

Remark. In any number field, one can factorise any $\alpha \in \mathcal{O}_L$ into a product of irreducibles by induction on $|N(\alpha)|$, but this factorisation is not in general unique. An idea due to Kummer is to measure the failure of unique factorisation by studying ideals $\mathfrak{a} \lhd \mathcal{O}_L$.

If $x_1, \dots, x_n \in \mathcal{O}_L$, we write $(x_1, \dots, x_n)$ for the ideal $\sum x_i \mathcal{O}_L$ generated by the x_i. We will consider products of ideals, rather than products of elements.

Definition. If $\mathfrak{a}, \mathfrak{b} \lhd \mathcal{O}_L$, define

$$\mathfrak{a} + \mathfrak{b} = \{x + y \mid x \in \mathfrak{a}, y \in \mathfrak{b}\}; \quad \mathfrak{a}\mathfrak{b} = \left\{\sum_i x_i y_i \,\middle|\, x_i \in \mathfrak{a}, y_i \in \mathfrak{b}\right\}$$

One can check that this is an ideal, and that products are associative.

Example. $(x_1, \dots, x_n)(y_1, \dots, y_m) = \big(\{x_i y_j \mid 1 \le i \le n, 1 \le j \le n\}\big)$. For instance, $(x)(y) = (xy)$, so the product of principal ideals is principal.

Example. Consider $\mathbb{Z}[\sqrt{5}] = \mathcal{O}_L$, and the ideals $\mathfrak{p}_1 = (3, 1 + 2\sqrt{5}), \mathfrak{p}_2 = (3, 1 - 2\sqrt{5})$. We obtain $\mathfrak{p}_1\mathfrak{p}_2 = (9, 3(1 - 2\sqrt{5}), 3(1 + 2\sqrt{5}), 21) = (3)$. So the ideal (3) factors as $\mathfrak{p}_1\mathfrak{p}_2$ in $\mathcal{O}_L$. Note that $37 = (1 + 2\sqrt{-5})(1 - 2\sqrt{-5})$, so $\mathbb{Z}[\sqrt{5}]$ is not a unique factorisation domain.

Recall that an ideal $\mathfrak{p} \lhd R$ is *prime* if ${}^{R}\!/\!_{\mathfrak{p}}$ is an integral domain, so $p \neq R$ and for all $x, y \in R$, $xy \in \mathfrak{p}$ implies $x \in \mathfrak{p}$ or $y \in \mathfrak{p}$. In this course, we will also define that a prime ideal is nonzero.

Lemma. If $\mathfrak{a} \lhd \mathcal{O}_K$, it contains an integer, and moreover, ${}^{\mathcal{O}_K}\!/\!_{\mathfrak{a}}$ is a finite abelian group.

Proof. Let $\alpha \in \mathfrak{a}, \alpha \neq 0$. Let $p_\alpha(x) = x^m + a_{m-1}x^{m-1} + \dots + a_0 \in \mathbb{Z}[x]$ be its minimal polynomial, and $a_0 \neq 0$. Then $a_0 = -\alpha(\alpha^{n-1} + a_{n-1}\alpha^{n-1} + \dots + a_2\alpha + a_1)$. But $a_0 \in \mathbb{Z}$, $\alpha \in \mathfrak{a}$, and the other factor lies in $\mathcal{O}_K$. So $a_0 \in \mathfrak{a}$ as $\mathfrak{a}$ is an ideal. Hence $a_0\mathcal{O}_K \subseteq \mathfrak{a}$, so ${}^{\mathcal{O}_K}\!/\!_{a_0\mathcal{O}_K}$ surjects onto ${}^{\mathcal{O}_K}\!/\!_{\mathfrak{a}}$. But for any integer d, ${}^{\mathcal{O}_K}\!/\!_{d\mathcal{O}_K} = {}^{\mathbb{Z}^n}\!/\!_{d\mathbb{Z}^n} = \left({}^{\mathbb{Z}}\!/\!_{d\mathbb{Z}}\right)^n$ is a finite set, so ${}^{\mathcal{O}_K}\!/\!_{\mathfrak{a}}$ is finite. □

Corollary. $\mathfrak{a} \simeq \mathbb{Z}^n$, as $\mathcal{O}_K \simeq \mathbb{Z}^n$ and the quotient is finite.

Therefore, nonzero ideals in $\mathcal{O}_K$ are isomorphic to $\mathbb{Z}^n$ as abelian groups.

Proposition. (i) $\mathcal{O}_K$ is an integral domain.

(ii) $\mathcal{O}_K$ is a Noetherian ring.

(iii) $\mathcal{O}_K$ is *integrally closed* in K (which is the field of fractions of $\mathcal{O}_K$): if $x \in K$ is integral over $\mathcal{O}_K$, it lies in $\mathcal{O}_K$.

(iv) Every (implicitly nonzero) prime ideal is maximal. We say that the *Krull dimension* of $\mathcal{O}_K$ is 1.

Remark. A ring with these four properties is called a *Dedekind domain*. Many of the results in this section hold for all Dedekind domains.

Proof. Part (i). $\mathcal{O}_K \subseteq K$, and K is a field.

Part (ii). We have shown that $\mathcal{O}_K \simeq \mathbb{Z}^n$, where $n = [K : \mathbb{Q}]$, so $\mathcal{O}_K$ is finitely generated as an abelian group, so is certainly finitely generated as a ring.

Part (iii). $\mathcal{O}_K$ is integral over $\mathbb{Z}$ by definition, so if x is integral over $\mathcal{O}_K$, it is integral over $\mathbb{Z}$. So x is an algebraic integer, so lies in $\mathcal{O}_K$.

Part (iv). If $\mathfrak{p}$ is a prime ideal, then by the previous lemma ${}^{\mathcal{O}_K}\!/\!_{\mathfrak{p}}$ is finite and an integral domain, as $\mathfrak{p}$ is prime. All finite integral domains are fields, hence $\mathfrak{p}$ is maximal. □

Example. Consider $R = \mathbb{C}[X, Y]$. Then (x) is prime but not maximal, since $(x) \subsetneq (x, y)$.

2.2. Unique factorisation of ideals

We aim to show that every ideal in $\mathcal{O}_K$ factors uniquely as a product of prime ideals.

Definition. $\mathfrak{b}$ divides $\mathfrak{a}$ if there exists an ideal $\mathfrak{c}$ such that $\mathfrak{a} = \mathfrak{b}\mathfrak{c}$. We write $\mathfrak{b} \mid \mathfrak{a}$.

Example. $(5, 1 + 2\sqrt{5}) \mid (3)$ in $\mathcal{O}_{\mathbb{Q}(\sqrt{-5})}$. $3\mathbb{Z} \mid 6\mathbb{Z}$ as $3\mathbb{Z} \cdot 2\mathbb{Z} = 6\mathbb{Z}$.

Note that $\mathfrak{b}\mathfrak{c} \subseteq \mathfrak{b}$, as $\mathfrak{b}$ is an ideal. So if $\mathfrak{b} \mid \mathfrak{a}$, then $\mathfrak{a} \subseteq \mathfrak{b}$. We will show the converse, that $\mathfrak{a} \subseteq \mathfrak{b}$ implies $\mathfrak{b} \mid \mathfrak{a}$. This allows us to prove results about division by using containment. Note that prime ideals are maximal, which allows us to use the containment relation.

Lemma. Let $\mathfrak{p}$ be a prime ideal in a ring R, and let $\mathfrak{a}, \mathfrak{b} \lhd R$ be ideals. Then if $\mathfrak{a}\mathfrak{b} \subseteq \mathfrak{p}$, either $\mathfrak{a} \subseteq \mathfrak{p}$ or $\mathfrak{b} \subseteq \mathfrak{p}$.

Proof. Otherwise, there exists $a \in \mathfrak{a} \setminus \mathfrak{p}$ and $b \in \mathfrak{b} \setminus \mathfrak{p}$, with $ab \in \mathfrak{p}$. But $\mathfrak{p}$ is prime giving a contradiction. □

Lemma. Let $\mathfrak{a} \trianglelefteq \mathcal{O}_K$ be a nonzero ideal. Then $\mathfrak{a}$ contains a product of prime ideals.

Proof. Otherwise, as $\mathcal{O}_K$ is Noetherian, there exists a ideal $\mathfrak{a}$ which is maximal with this property. In particular, $\mathfrak{a}$ is not prime. So there exists $x, y \in \mathcal{O}_K$ with x or y not in $\mathfrak{a}$ but $xy \in \mathfrak{a}$. So $\mathfrak{a} \subsetneq \mathfrak{a} + (x)$. But then, $\mathfrak{a} + (x)$ contains a product of prime ideals $\mathfrak{p}_1, \dots, \mathfrak{p}_r$ with $\mathfrak{p}_1 \dots \mathfrak{p}_r \subseteq \mathfrak{a} + (x)$. Similarly, there exist prime ideals $\mathfrak{q}_1, \dots \mathfrak{q}_s$ such that $\mathfrak{q}_1 \dots \mathfrak{q}_s \subseteq \mathfrak{a} + (y)$. Then,

$$\mathfrak{p}_1 \dots \mathfrak{p}_r \mathfrak{q}_1 \dots \mathfrak{q}_s \subseteq (\mathfrak{a} + (x))(\mathfrak{a} + (y)) = \mathfrak{a} + (xy)$$

But $xy \in \mathfrak{a}$, giving a contradiction. □

The main proof will use the idea that we can formally introduce the group of fractions of the commutative monoid of ideals. The object $\{y \in K \mid y\mathfrak{a} \subseteq \mathcal{O}_K\}$ will represent the inverse of $\mathfrak{a}$.

Lemma. (i) Let $0 \neq \mathfrak{a} \trianglelefteq \mathcal{O}_K$ be an ideal. If $x \in K$ has the property that $x\mathfrak{a} \subseteq \mathfrak{a}$, then $x \in \mathcal{O}_K$.

(ii) Let $0 \neq \mathfrak{a} \lhd \mathcal{O}_K$ be a proper ideal. Then, $\mathcal{O}_K \subseteq \{y \in K \mid y\mathfrak{a} \subseteq \mathcal{O}_K\}$ contains elements which are not in $\mathcal{O}_K$. Equivalently, ${}^{\{y \in K \mid y\mathfrak{a} \subseteq \mathcal{O}_K\}}\!/_{\mathcal{O}_K} \neq \{1\}$ as abelian groups.

Example. Let $\mathcal{O}_K = \mathbb{Z}$ and $\mathfrak{a} = 3\mathbb{Z}$. Then, part (i) shows that if $\frac{a}{b} \cdot 3 \in 3\mathbb{Z}$, then $\frac{a}{b} \in \mathbb{Z}$. Part (ii) shows that if $\frac{a}{b} \cdot 3 \in \mathbb{Z}$ then $\frac{a}{b} \in \frac{1}{3}\mathbb{Z}$; for instance, if $\frac{a}{b} = \frac{1}{3}$, we have ${}^{\frac{1}{3}\mathbb{Z}}\!/_{\mathbb{Z}} = {}^{\mathbb{Z}}\!/_{3\mathbb{Z}} \neq \{1\}$.

Proof. Part (i). $\mathfrak{a} \subseteq \mathcal{O}_K$ is finitely generated as an abelian group, as it is isomorphic to $\mathbb{Z}^n$. Let $\alpha_1, \dots, \alpha_n$ be a $\mathbb{Z}$-basis of $\mathfrak{a}$. Consider $m_x : \mathfrak{a} \to \mathfrak{a}$ given by multiplication by $x \in K$. We write $x\alpha_i = \sum a_{ij}\alpha_j$, where by assumption, a_{ij} are integers. Hence,

$$(xI - A)\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = 0$$

where $A = (a_{ij})$. So $\det(xI - A) = 0$, so x is integral over $\mathbb{Z}$; that is, $x \in \mathcal{O}_K$.

Part (ii). If this holds for $\mathfrak{a}$, it certainly holds for all ideals $\mathfrak{a}' \subseteq \mathfrak{a}$. So without loss of generality, let $\mathfrak{a}$ be maximal, so $\mathfrak{a} = \mathfrak{p}$ is a prime ideal. Let $\alpha \in \mathfrak{p}$ be nonzero. By the previous lemma, there exist prime ideals $\mathfrak{p}_1, \dots, \mathfrak{p}_r$ such that $\mathfrak{p}_1 \dots \mathfrak{p}_r \subseteq (\alpha) \subseteq \mathfrak{p}$. Suppose that r is minimal. By the first lemma in this subsection, there exists i such that $\mathfrak{p}_i \subseteq \mathfrak{p}$, and without loss of generality $i = 1$. So $\mathfrak{p}_1 \subseteq \mathfrak{p}$. But $\mathfrak{p}_1$ is maximal, so $\mathfrak{p}_1 = \mathfrak{p}$.

Since r is minimal, $\mathfrak{p}_2 \dots \mathfrak{p}_r \subsetneq (\alpha)$. Fix $\beta \in \mathfrak{p}_2 \dots \mathfrak{p}_r \setminus (\alpha)$. Then $\beta\mathfrak{p} \subseteq \mathfrak{p}(\mathfrak{p}_2 \dots \mathfrak{p}_r) \subseteq (\alpha)$, but $\beta \in (\alpha)$. So, dividing by α, we obtain $\frac{\beta}{\alpha}\mathfrak{p} \subseteq (1) = \mathcal{O}_K$, but $\frac{\beta}{\alpha} \notin \mathcal{O}_K$. □

Definition. A *fractional ideal* is an $\mathcal{O}_K$-module X such that $X \subseteq K$ and X is finitely generated.

$X = \{x \in K \mid x\mathfrak{a} \subseteq \mathcal{O}_K\}$ is an $\mathcal{O}_K$-module. If $\alpha \in \mathfrak{a} \setminus \{0\}$, then $\alpha X \subseteq \mathcal{O}_K = \mathbb{Z}^n$ where $n = [K : \mathbb{Q}]$. Multiplication by α is an isomorphism $X \to \alpha X$, and submodules of $\mathbb{Z}^n$ are finitely generated abelian groups, so X is finitely generated as an abelian group, hence as an $\mathcal{O}_K$-module. Hence X is a fractional ideal.

Lemma. $\mathfrak{q} \subseteq K$ is a fractional ideal if and only if there exists a nonzero constant $c \in K$ such that $c\mathfrak{q}$ is an ideal in $\mathcal{O}_K$.

Proof. Suppose $c\mathfrak{q}$ is an ideal. Then $\mathfrak{q} \subseteq K$, and multiplication by c is an isomorphism $\mathfrak{q} \to c\mathfrak{q}$ as $\mathcal{O}_K$-modules, so it is finitely generated as $\mathfrak{q}$ is.

Suppose $\mathfrak{q}$ is a fractional ideal. Then, $x_1, \dots, x_r$ generate $\mathfrak{q}$ as an $\mathcal{O}_K$-module. But $x_i \in K$ so $x_i = \frac{y_i}{n_i}$ where $y_i \in \mathcal{O}_K$, $n_i \in \mathbb{Z}$. Let c be the least common multiple of the n_i, and then $cq \subseteq \mathcal{O}_K$, and is a submodule of $\mathcal{O}_K$, and hence is an ideal. □

Corollary. $\mathfrak{q}$ is isomorphic to $\mathbb{Z}^n$ as an abelian group.

Proof. We have shown that all nonzero ideals in $\mathcal{O}_K$ are isomorphic to $\mathbb{Z}^n$ as abelian groups, where $n = [K : \mathbb{Q}]$, and multiplication by c is an isomorphism $\mathfrak{q} \to c\mathfrak{q}$. □

Ideals are sometimes called *integral ideals* to distinguish from fractional ideals. One can define multiplication of fractional ideals in the same way that we defined it for integral ideals.

Definition. A fractional ideal $\mathfrak{q}$ is *invertible* if there exists a fractional ideal $\mathfrak{r}$ such that $\mathfrak{q}\mathfrak{r} = (1) = \mathcal{O}_K$.

Proposition. Every nonzero fractional ideal $\mathfrak{q}$ is invertible, and its inverse is

$$\mathfrak{q}^{-1} = \{x \in K \mid x\mathfrak{q} \subseteq \mathcal{O}_K\}$$

Remark. $\mathfrak{q} = \frac{1}{n}\mathfrak{a}, \mathfrak{r} = \frac{1}{m}\mathfrak{b}$ where $\mathfrak{a}, \mathfrak{b}$ are integral ideals in $\mathcal{O}_K$, and $n, m \in K^\star$. Then $\mathfrak{q}\mathfrak{r} = 1$ if and only if $\mathfrak{a}\mathfrak{b} = (nm)$. Therefore, the proposition is equivalent to the statement that for every $\mathfrak{a} \trianglelefteq \mathcal{O}_K$, there exists an ideal $\mathfrak{b} \trianglelefteq \mathcal{O}_K$ such that $\mathfrak{a}\mathfrak{b}$ is principal.

Proof. $\mathfrak{q}$ is invertible if and only if $\mathfrak{a}$ is invertible, where $n\mathfrak{q} = \mathfrak{a}$ as above. So, without loss of generality, let $\mathfrak{q}$ be an integral ideal. If the proposition is false, there exists some integral ideal in $\mathcal{O}_K$. As $\mathcal{O}_K$ is Noetherian, there exists a maximal such ideal $\mathfrak{a} \neq \mathcal{O}_K$. So every ideal $\mathfrak{a}' \supsetneq \mathfrak{a}$ is invertible. Let $\mathfrak{b} = \{x \in K \mid x\mathfrak{a} \subseteq \mathcal{O}_K\}$, which is a fractional ideal. $\mathcal{O}_K \subseteq \mathfrak{b}$ hence $\mathfrak{a} \subseteq \mathfrak{a}\mathfrak{b}$. If $\mathfrak{a} = \mathfrak{a}\mathfrak{b}$, then part (i) of a previous lemma implies that $\mathfrak{b} \subseteq \mathcal{O}_K$. Part (ii) of the same lemma implies $\mathfrak{b} \setminus \mathcal{O}_K \neq \varnothing$, which is a contradiction. So $\mathfrak{a} \subsetneq \mathfrak{a}\mathfrak{b} \subsetneq \mathcal{O}_K$. Then $\mathfrak{a}\mathfrak{b}$ is invertible by assumption, so $\mathfrak{a}$ is invertible, giving a contradiction. Finally, $\mathfrak{q}^{-1} \subseteq \{x \in K \mid x\mathfrak{q} \subseteq \mathcal{O}_K\} = X$, so $\mathfrak{q}\mathfrak{q}^{-1} = \mathcal{O}_K \subseteq \mathfrak{q}X \subseteq \mathcal{O}_K$, so we have equality: $\mathfrak{q}^{-1} = X$. □

Corollary. Let $\mathfrak{a}, \mathfrak{b}, \mathfrak{c} \lhd \mathcal{O}_K$ be integral ideals, and let $\mathfrak{c} \neq (0)$. Then,

(i) $\mathfrak{b} \subseteq \mathfrak{a} \iff \mathfrak{b}\mathfrak{c} \subseteq \mathfrak{a}\mathfrak{c}$;

(ii) $\mathfrak{a} \mid \mathfrak{b} \iff \mathfrak{a}\mathfrak{c} \mid \mathfrak{b}\mathfrak{c}$;

(iii) $\mathfrak{a} \mid \mathfrak{b} \iff \mathfrak{b} \subseteq \mathfrak{a}$.

Proof. The forward direction of parts (i) and (ii) are clear; the backward direction follows from multiplication by $\mathfrak{c}^{-1}$. The forward direction of part (iii) has already been seen. Now, suppose $\mathfrak{b} \subseteq \mathfrak{a}$. By the proposition above, there exists $\mathfrak{c}$ such that $\mathfrak{a}\mathfrak{c} = (\alpha)$ is principal. Then, $\mathfrak{b} \subseteq \mathfrak{a}$ if and only if $\mathfrak{b}\mathfrak{c} \subseteq (\alpha)$ by part (i). $\mathfrak{a} \mid \mathfrak{b}$ if and only if $(\alpha) \mid \mathfrak{b}\mathfrak{c}$ by part (ii). But if $\mathfrak{b}\mathfrak{c}$ is generated by $\beta_1, \dots, \beta_r$, $\mathfrak{b}\mathfrak{c} \subseteq (\alpha)$ means that each β_i is divisible by α. More precisely, $\beta_i = \beta_i'\alpha$ for some $\beta_i' \in \mathcal{O}_K$. So $(\beta_1, \dots, \beta_r) = (\beta_1', \dots, \beta_r')(\alpha)$ proving part (iii). □

Remark. Part (iii) is straightforward if $\mathfrak{a}$ is principal, and invertibility via fractional ideals allows us to reduce to this case.

Theorem. Let $\mathfrak{a} \lhd \mathcal{O}_K$ be a nonzero ideal. Then $\mathfrak{a}$ can be written uniquely as a product of prime ideals.

Proof. If $\mathfrak{a}$ is not prime, it is not maximal. Let $\mathfrak{b} \supsetneq \mathfrak{a}$ be an ideal in $\mathcal{O}_K$. Then $\mathfrak{a} = \mathfrak{b}\mathfrak{c}$ for some ideal $\mathfrak{c}$ containing $\mathfrak{a}$ by part (iii) of the previous corollary. We continue factoring in this way. As the ring is Noetherian, this process will always terminate, as we produce an ascending chain.

For uniqueness, we have shown that $\mathfrak{p} \mid \mathfrak{a}\mathfrak{b}$ implies $\mathfrak{p} \mid \mathfrak{a}$ or $\mathfrak{p} \mid \mathfrak{b}$. So if $\mathfrak{p}_1 \dots \mathfrak{p}_r = \mathfrak{q}_1 \dots \mathfrak{q}_s$ with $\mathfrak{p}_i, \mathfrak{q}_i$ prime, we have $\mathfrak{p}_1 \mid \mathfrak{q}_i$ for some i. So let $i = 1$ without loss of generality, so $\mathfrak{q}_1 \subseteq \mathfrak{p}_1$. But $\mathfrak{q}_1$ is maximal, so $\mathfrak{q}_1 = \mathfrak{p}_1$. Multiply by $\mathfrak{p}_1^{-1}$ to obtain $\mathfrak{p}_2 \dots \mathfrak{p}_r = \mathfrak{q}_2 \dots \mathfrak{q}_s$, then by induction, the $\mathfrak{p}_i$ and $\mathfrak{q}_i$ match. □

Corollary. The nonzero fractional ideals form a group I_K under multiplication. I_K is the free abelian group generated by the prime ideals $\mathfrak{p} \lhd \mathcal{O}_K$. In other words, any $\mathfrak{q} \in I_K$ can be written uniquely as a product of prime ideals and their inverses. $\mathfrak{q} \in I_K$ is an integral ideal if and only if all of the exponents are nonnegative.

Proof. Follows from the previous theorem after writing $\mathfrak{q} = \mathfrak{a}\mathfrak{b}^{-1}$ where $\mathfrak{a}, \mathfrak{b} \trianglelefteq \mathcal{O}_K$. □

2.3. Class group

Observe that we have a map $K^\star \to I_K$ mapping x to the principal ideal (x). This map is a group homomorphism, as $\alpha\beta \mapsto (\alpha)(\beta)$. Its kernel is the set of $\alpha \in K^\star$ such that $(\alpha) = (1) = \mathcal{O}_K$, which is the set $\mathcal{O}_K^\star$ of invertible elements of $\mathcal{O}_K$. The image is the set of principal ideals P_K.

Definition. The *class group* of a number field K is $\mathrm{Cl}_K = {}^{I_K}\!/\!_{P_K}$, the cokernel of the map $K^\star \to I_K$.

If $\mathfrak{a} \in I_K$, we write $[\mathfrak{a}]$ for its equivalence class in the class group, so $[\mathfrak{a}] = [\mathfrak{b}]$ if and only if there exists $\gamma \in K^\star$ such that $\gamma\mathfrak{a} = \mathfrak{b}$.

Theorem. The following are equivalent.

(i) $\mathcal{O}_K$ is a principal ideal domain;

(ii) $\mathcal{O}_K$ is a unique factorisation domain;

(iii) Cl_K is trivial.

Proof. (i) holds if and only if (iii) holds by definition. (i) implies (ii) is a general fact from IB Groups, Rings and Modules. The proof that (ii) implies (i) remains. Let $\mathfrak{p}$ be a prime ideal in $\mathcal{O}_K$, and $x \in \mathfrak{p}$ a nonzero element of this ideal. We can factorise x into irreducibles $x = \alpha_1 \dots \alpha_r$ uniquely by assumption. As $\mathfrak{p}$ is prime, some α_i lies in $\mathfrak{p}$. Then $(\alpha_i) \subseteq \mathfrak{p}$, and as $\mathcal{O}_K$ is a unique factorisation domain and α_i is irreducible, (α_i) is prime. But prime ideals are maximal, so $(\alpha_i) = \mathfrak{p}$ as required. □

The following sequence is exact.

$$\begin{tikzcd} 1 \arrow[r] & \mathcal{O}_K^\star \arrow[r] & K^\star \arrow[r] & I_K \arrow[r] & \mathrm{Cl}_K \arrow[r] & 1 \end{tikzcd}$$

We can now state the main theorems of the course, which are:

(i) the class group is finite;

(ii) $\mathcal{O}_K^\star$ is the direct product of the roots of unity in K with $\mathbb{Z}^{r+s-1}$.

Example. $(3, 1+2\sqrt{5})(3, 1-2\sqrt{5}) = (3)$, so $(3, 1+2\sqrt{5})$ and $(3, 1-2\sqrt{5})$ are inverse in the class group.

Example. Let $[L : \mathbb{Q}] = 2$, so $L = \mathbb{Q}(\sqrt{d})$ for $d \in \mathbb{Z}$, and $d \not\equiv 1 \bmod 4$. Let $\mathfrak{a} \trianglelefteq \mathcal{O}_L$, so $\mathfrak{a} \simeq \mathbb{Z}^2$ giving $\mathfrak{a} = (\alpha, \beta)$ as an $\mathcal{O}_L$-module. We can always assume $\beta \in \mathbb{Z}$. Indeed, write $\alpha = a + b\sqrt{d}$ and $\beta = a' + b'\sqrt{d}$. Assume $|a| + |a'|$ is minimal, so without loss of generality $a \geq a' \geq 0$, and if $a' \neq 0$, $\alpha - \beta, \beta$ has smaller $|a| + |a'|$.

Example. In a quadratic field $\mathfrak{a} = (\alpha, b)$ where $b \in \mathbb{Z}$. Then $(b, \alpha)(b, \overline{\alpha})$ is principal.

$$\mathfrak{a}\overline{\mathfrak{a}} = (b^2, b\alpha, b\overline{\alpha}, \alpha\overline{\alpha}) = (b^2, b\alpha, b\underbrace{(\alpha + \overline{\alpha})}_{\mathrm{Tr}(\alpha)}, N(\alpha)) = (b\alpha, c)$$

where $c = \gcd(b^2, \operatorname{Tr}(\alpha), N(\alpha))$. Let $x = \frac{b\alpha}{c} \in L^\star$. $\operatorname{Tr}(x) = \frac{b\operatorname{Tr}(\alpha)}{c} \in \mathbb{Z}$, and $N(c) = N\left(\frac{b\alpha}{c}\right) = \frac{b^2 N(\alpha)}{c^2} = \frac{b^2}{c}\frac{N(\alpha)}{c} \in \mathbb{Z}$, so $x \in \mathcal{O}_L$, giving $c \mid b\alpha$, so $\mathfrak{a}\overline{\mathfrak{a}} = (c)$. In particular, $(b,\alpha), (b,\overline{\alpha})$ are inverse in the class group.

2.4. Norms of ideals

Definition. Let L be a number field, and let $[L : \mathbb{Q}] = n$. Let $\mathfrak{a} \trianglelefteq \mathcal{O}_L$ be a nonzero ideal. The *norm* of $\mathfrak{a}$ is $\left|\mathcal{O}_L/\mathfrak{a}\right|$.

By Lagrange's theorem, $N(\mathfrak{a}) \cdot 1 = 0$ in $\mathcal{O}_L/\mathfrak{a}$. Hence $N(\mathfrak{a}) \in \mathfrak{a} \cap \mathbb{Z}$.

Example. Let p be a prime. $N((p)) = \left|\mathbb{Z}^n/(p\mathbb{Z})^n\right| = p^n$.

Proposition. Let $\mathfrak{a}, \mathfrak{b} \trianglelefteq \mathcal{O}_L$ be nonzero ideals. Then, $N(\mathfrak{a}\mathfrak{b}) = N(\mathfrak{a})N(\mathfrak{b})$.

Remark. By unique factorisation of ideals, it suffices to show that

$$N(\mathfrak{p}_1^{a_1} \dots \mathfrak{p}_n^{a_n}) = N(\mathfrak{p}_1)^{a_1} \dots N(\mathfrak{p}_n)^{a_n}$$

for $\mathfrak{p}_i$ distinct prime ideals. To show this, we need that

(i) $\mathcal{O}_L/\mathfrak{p}_1^{a_1} \dots \mathfrak{p}_n^{a_n} \simeq \mathcal{O}_L/\mathfrak{p}_1^{a_1} \dots \mathcal{O}_L/\mathfrak{p}_n^{a_n}$ by the Chinese remainder theorem.

(ii) $\left|\mathcal{O}_L/\mathfrak{p}^e\right| = \left|\mathcal{O}_L/\mathfrak{p}\right| \cdot \left|\mathfrak{p}/\mathfrak{p}^2\right| \cdots \left|\mathfrak{p}^{e-1}/\mathfrak{p}^e\right|$ which is a general fact, and this is equal to $\left|\mathcal{O}_L/\mathfrak{p}\right|^e$ as $\mathfrak{p}^a/\mathfrak{p}^{a+1}$ is a one-dimensional vector space over the field $\mathcal{O}_L/\mathfrak{p}$. This fact is specific to number fields (or more generally, Dedekind domains). For a counterexample, consider $\mathbb{F}_p[X,Y]$ and $\mathfrak{p} = (x,y)$.

The following proof uses the above approach obscurely but quickly.

Proof. By unique factorisation it suffices to show the result for $\mathfrak{b} = \mathfrak{p}$ where $\mathfrak{p}$ is prime. $\mathfrak{a} \neq \mathfrak{a}\mathfrak{p}$ by unique factorisation, so let $\alpha \in \mathfrak{a} \setminus \mathfrak{a}\mathfrak{p}$. We claim that the homomorphism of abelian groups $\mathcal{O}_L/\mathfrak{p} \to \mathfrak{a}/\mathfrak{a}\mathfrak{p}$ mapping $x \mapsto \alpha x$ is an isomorphism. Then,

$$\mathcal{O}_L/\mathfrak{a} \simeq \left(\mathcal{O}_L/\mathfrak{a}\mathfrak{p}\right)/\left(\mathfrak{a}/\mathfrak{a}\mathfrak{p}\right)$$

so

$$N(\mathfrak{a}) = \left|\mathcal{O}_L/\mathfrak{a}\right| = \frac{N(\mathfrak{a}\mathfrak{p})}{\left|\mathfrak{a}/\mathfrak{a}\mathfrak{p}\right|}$$

but $\left|\mathfrak{a}/\mathfrak{a}\mathfrak{p}\right| = \left|\mathcal{O}_L/\mathfrak{p}\right| = N(\mathfrak{p})$ by the claim, proving the proposition. We now prove the claim.

We show the homomorphism is injective. $(\alpha) \subseteq \mathfrak{a}$ so $(\alpha) = \mathfrak{a}\mathfrak{c}$ for some $\mathfrak{c} \vartriangleleft \mathcal{O}_L$. Suppose x has $\alpha x \in \mathfrak{a}\mathfrak{p}$, so $x + \mathfrak{p}$ is in the kernel. Then, $x\mathfrak{a}\mathfrak{c} \subseteq \mathfrak{a}\mathfrak{p}$. Dividing by $\mathfrak{a}$, $x\mathfrak{c} \subseteq \mathfrak{p}$. But $\mathfrak{p}$ is prime, so $x \in p$ or $\mathfrak{c} \subseteq \mathfrak{p}$. But $\mathfrak{c} \subseteq \mathfrak{p}$ implies $\alpha \in \mathfrak{a}\mathfrak{p}$, contradicting our choice of α. So $x \in \mathfrak{p}$, so the map is injective as required.

We show the homomorphism is surjective. We want to show $(\alpha) + \mathfrak{a}\mathfrak{p} = \mathfrak{a}$. We know that $\mathfrak{a}\mathfrak{p} \subsetneq (\alpha) + \mathfrak{a}\mathfrak{p} \subseteq \mathfrak{a}$. Multiplying by $\mathfrak{a}^{-1}$, we obtain

$$\mathfrak{p} \subsetneq ((\alpha) + \mathfrak{a}\mathfrak{p})\mathfrak{a}^{-1} \subseteq \mathcal{O}_L$$

But $\mathfrak{p}$ is a prime and hence maximal. Therefore, $((\alpha) + \mathfrak{a}\mathfrak{p})\mathfrak{a}^{-1} = \mathcal{O}_L$, so $(\alpha) + p = \mathfrak{a}$, so the map is surjective. □

Lemma. Let $M \subseteq \mathbb{Z}^n$ be a subgroup. Then $M \simeq \mathbb{Z}^r$ for some $0 \leq r \leq n$. Suppose further that $r = n$. Let $e_1, \dots, e_n$ be a basis of $\mathbb{Z}^n$ and $v_1, \dots, v_n$ be a basis of M over $\mathbb{Z}$. Then, $\left|\mathbb{Z}^n/M\right| = \det A$ where $A = (a_{ij})$ and $v_j = \sum a_{ij} e_i$.

Proof. We can choose a basis $v_1, \dots, v_n$ of M such that A is upper triangular. Then, $|\det A| = |a_{11} \dots a_{nn}|$. □

Lemma. Let $\mathfrak{a} \lhd \mathcal{O}_L$ be a nonzero ideal, and $n = [L : \mathbb{Q}]$. Then,

(i) There exist $\alpha_1, \dots, \alpha_n \in \mathfrak{a}$ such that $\mathfrak{a} = \left\{\sum_{i=1}^n r_i \alpha_i \mid r_i \in \mathbb{Z}\right\}$, and $\alpha_1, \dots, \alpha_n$ are a basis of $L/\mathbb{Q}$.

(ii) For any such $\alpha_1, \dots, \alpha_n \in \mathfrak{a}$, $\Delta(\alpha_1, \dots, \alpha_n) = N(\mathfrak{a})^2 D_L$ where D_L is the discriminant of L, and where $\Delta(\alpha_1, \dots, \alpha_n) = \det \operatorname{Tr}(\alpha_i \alpha_j) = \left(\det(\sigma_i \alpha_j)\right)^2$.

Proof. Part (i). The result holds for $\mathcal{O}_L$, and if $d \in \mathfrak{a}$ is an integer, such as $d = N(\mathfrak{a})$, then $d\mathcal{O}_L \subseteq \mathfrak{a} \subseteq \mathcal{O}_L$, so as abelian groups, $(d\mathbb{Z})^n \subseteq \mathfrak{a} \subseteq \mathbb{Z}^n$, so $\mathfrak{a} \simeq \mathbb{Z}^n$.

Part (ii). Let $\alpha_1', \dots, \alpha_n'$ be an integral basis of $\mathcal{O}_L$. Let A be the change of basis matrix from $\alpha_1, \dots, \alpha_n$ to $\alpha_1', \dots, \alpha_n'$. Then $\Delta(\alpha_1, \dots, \alpha_n) = (\det A)^2 \Delta(\alpha_1', \dots, \alpha_n') = \left|\mathcal{O}_L/\mathfrak{a}\right|^2 D_L$ by the lemma. □

Corollary. If $\alpha_1, \dots, \alpha_n$ generating $\mathfrak{a}$ as a $\mathbb{Z}$-module has $\Delta(\alpha_1, \dots, \alpha_n)$ square-free, then $\mathfrak{a} = \mathcal{O}_L$ and D_L is square-free. In particular, if $L = \mathbb{Q}(\alpha)$ and $\alpha \in \mathcal{O}_L$ where the discriminant $\operatorname{disc}(\alpha) = \Delta(1, \alpha, \dots, \alpha^{n-1})$ is square-free, then $\mathbb{Z}[\alpha] = \mathcal{O}_L$. More generally, if $\alpha \in \mathcal{O}_L$ and $L = \mathbb{Q}(\alpha)$, and $d \in \mathbb{Z}$ is a maximal integer such that $d^2 \mid \operatorname{disc}(\alpha)$, then $\mathbb{Z}[\alpha] \subseteq \mathcal{O}_L \subseteq \frac{1}{d}\mathbb{Z}[\alpha]$.

Lemma. Let $\alpha \in \mathcal{O}_L$ be a nonzero algebraic integer. Then $N((\alpha)) = |N_{L/\mathbb{Q}}(\alpha)|$.

Proof. Let $\alpha_1, \dots, \alpha_n$ be an integral basis of $\mathcal{O}_L$. Consider

$$\begin{aligned}\Delta(\alpha_1\alpha, \dots, \alpha_n\alpha) &= \big(\det(\sigma_i(\alpha_j\alpha))\big)^2 \\ &= \big(\det((\sigma_i\alpha_j)(\sigma_i\alpha))\big)^2 \\ &= \left(\prod_{i=1}^n \sigma_i(\alpha)\cdot\det(\sigma_i\alpha_j)\right)^2 \\ &= N(\alpha)^2\Delta(\alpha_1, \dots, \alpha_n) \\ &= N(\alpha)^2 D_L\end{aligned}$$

But $\alpha_1\alpha, \dots, \alpha_n\alpha$ is a basis of (α), hence this is equal to $N((\alpha))^2 D_L$. So $N((\alpha))^2 = N_{L/\mathbb{Q}}(\alpha)^2$, but $N((\alpha)) > 0$, giving the result as required. □

2.5. Prime ideals

Lemma. Let $\mathfrak{p} \lhd \mathcal{O}_L$ be a prime ideal. Then there exists a unique prime $p \in \mathbb{Z}$ such that $\mathfrak{p} \mid (p) = p\mathcal{O}_L$. Moreover, $N(\mathfrak{p}) = p^f$ for some integer $1 \leq f \leq n = [L : \mathbb{Q}]$.

Proof. $\mathfrak{p} \cap \mathbb{Z}$ is an ideal in $\mathbb{Z}$, hence principal. So for some $p \in \mathbb{Z}$, $\mathfrak{p} \cap \mathbb{Z} = p\mathbb{Z}$; we claim p is prime. If $p = ab$ with $a, b \in \mathbb{Z}$, then as $p \in \mathfrak{p}$, a or b lie in $\mathfrak{p} \cap \mathbb{Z}$, so a or b lie in $p\mathbb{Z}$, so $p \mid a$ or $p \mid b$. By factorisation of ideals, $(p) = \mathfrak{p}\mathfrak{a}$ for some $\mathfrak{a} \trianglelefteq \mathcal{O}_L$. Taking norms, $N((p)) = N(\mathfrak{p})N(\mathfrak{a})$. But $N((p)) = p^n$, so $N(\mathfrak{p}) = p^f$ for $1 \leq f \leq n$. □

Remark. Every prime ideal in $\mathcal{O}_L$ is a factor of $(p) \lhd \mathbb{Z}$ where p is a prime. Hence, we can factorise (p) as $\mathfrak{p}_1^{e_1} \dots \mathfrak{p}_r^{e_r}$ for each prime $p \in \mathbb{Z}$ to identify all prime ideals in $\mathcal{O}_L$.

Let $p \in \mathbb{Z}$ be a prime. Consider the map $q : \mathcal{O}_L \to {}^{\mathcal{O}_L}\!/_{p\mathcal{O}_L}$, which is a surjection. By the isomorphism theorem, there is a bijection $I \mapsto q^{-1}(I)$ with inverse $J \mapsto {}^{J}\!/_{(p)}$ between the set of ideals in ${}^{\mathcal{O}_L}\!/_{p\mathcal{O}_L}$ and ideals of $\mathcal{O}_L$ containing $p\mathcal{O}_L$, or equivalently, ideals $\mathfrak{p} \lhd \mathcal{O}_L$ with $\mathfrak{p} \mid (p)$. The bijection maps prime ideals to prime ideals.

Under certain assumptions, we can determine the prime ideals in ${}^{\mathcal{O}_L}\!/_{(p)}$ exactly.

Theorem (Dedekind's criteria)**.** Let $\alpha \in \mathcal{O}_L$ have minimal polynomial $g(x) \in \mathbb{Z}[x]$. Suppose that $\mathbb{Z}[\alpha] \subseteq \mathcal{O}_L$ has finite index $\left|{}^{\mathcal{O}_L}\!/_{\mathbb{Z}[\alpha]}\right|$ not divisible by p. Let $\overline{g}(x) = g(x) \bmod p \in \mathbb{F}_p[x]$. Let $\overline{g}(x) = \overline{\varphi}_1^{e_1} \dots \overline{\varphi}_r^{e_r}$ be the factorisation of $g(x)$ into irreducibles in $\mathbb{F}_p[x]$. Then $p\mathcal{O}_L = (p) = \mathfrak{p}_1^{e_1} \dots \mathfrak{p}_r^{e_r}$ where $\mathfrak{p}_i = (p, \varphi_i(\alpha))$ is the prime ideal in $\mathcal{O}_L$ where we choose any monic polynomial $\varphi_i(x) \in \mathbb{Z}[x]$ which has reduction mod p equal to $\overline{\varphi}_i(x)$.

Proof. First, we show that each factor $\overline{\varphi}_i$ defines a prime ideal in ${}^{\mathbb{Z}[\alpha]}\!/_{p\mathbb{Z}[\alpha]}$. We will then relate this to prime ideals in ${}^{\mathcal{O}_L}\!/_{p\mathcal{O}_L}$. We have a surjective ring homomorphism $\mathbb{Z}[x] \to {}^{\mathbb{F}_p[x]}\!/_{\overline{\varphi}_i}$.

We claim that the kernel of this homomorphism is the ideal generated by p, φ_i. We can factor the map as $\mathbb{Z}[\alpha] \to \mathbb{F}_p[x] \to {}^{\mathbb{F}_p[x]}\!/\!_{\overline{\varphi}_i}$. It is clear that p, φ_i lie in the kernel. If $f \mapsto 0$, then $\overline{\varphi}_i \mid \overline{f}$ so there exists $\overline{h} \in \mathbb{F}_p[x]$ such that $\overline{f} = \overline{\varphi}_i \overline{h}$, so $f = \varphi_i h + ps$ for any lift h of $\overline{h}$ of the same degree. So the kernel is precisely (p, φ_i).

We can alternatively factor the map as $\mathbb{Z}[\alpha] \to {}^{\mathbb{Z}[x]}\!/\!_{g(x)\mathbb{Z}[x]} \to {}^{\mathbb{F}_p[x]}\!/\!_{\overline{\varphi}_i}$. We claim that the kernel of the map $\mathbb{Z}[\alpha] \to \mathbb{F}_p[\alpha] = {}^{\mathbb{F}_p[x]}\!/\!_{\overline{\varphi}_i}$ is the ideal $\mathfrak{q}_i \lhd \mathbb{Z}[\alpha]$ generated by $p, \varphi_i(\alpha)$. The proof of this claim is left as an exercise. Therefore, ${}^{\mathbb{Z}[\alpha]}\!/\!_{\mathfrak{q}_i} \simeq {}^{\mathbb{F}_p[x]}\!/\!_{\overline{\varphi}_i(x)}$. But $\overline{\varphi}_i(x)$ is irreducible by hypothesis, so ${}^{\mathbb{F}_p[x]}\!/\!_{\overline{\varphi}_i(x)}$ is a field, hence $\mathfrak{q}_i$ is a prime ideal. Therefore, ${}^{\mathbb{F}_p[x]}\!/\!_{\overline{\varphi}_i} \simeq \mathbb{F}_q$ where $q = p^{f_i}$ is some power of p. In particular, $\left|{}^{\mathbb{Z}[\alpha]}\!/\!_{\mathfrak{q}_i}\right| = \left|{}^{\mathbb{F}_p[x]}\!/\!_{\overline{\varphi}_i(x)}\right| = p^{f_i}$ where $f_i = \deg \overline{\varphi}_i$.

Now, if $\mathbb{Z}[\alpha] = \mathcal{O}_L$ the first part implies that $\mathfrak{p}_i = \mathfrak{q}_i$ is a prime ideal containing p, and $N(\mathfrak{p}_i) = p^{f_i}$. Suppose $p \nmid \left|{}^{\mathcal{O}_L}\!/\!_{\mathbb{Z}[\alpha]}\right|$. We claim that the inclusion map defines an isomorphism $\iota \colon {}^{\mathbb{Z}[\alpha]}\!/\!_{p\mathbb{Z}[\alpha]} \to {}^{\mathcal{O}_L}\!/\!_{p\mathcal{O}_L}$. This implies that there is a bijection between ideals of ${}^{\mathbb{Z}[\alpha]}\!/\!_{p\mathbb{Z}[\alpha]}$ and ideals of ${}^{\mathcal{O}_L}\!/\!_{p\mathcal{O}_L}$. Hence, there is a bijection between ideals of $\mathbb{Z}[\alpha]$ containing p and ideals of $\mathcal{O}_L$ containing p, where this bijection maps an ideal $(p, y) \trianglelefteq \mathbb{Z}[\alpha]$ to $\mathfrak{p} \trianglelefteq \mathcal{O}_L$ generated by the same elements under the inclusion map. In other words, it maps an ideal $\mathfrak{q}$ to $\mathfrak{q}\mathcal{O}_L$. The inverse bijection maps $\mathfrak{p}$ to $\mathfrak{p} \cap \mathbb{Z}[\alpha]$. Moreover, ${}^{\mathcal{O}_L}\!/\!_{\mathfrak{p}} \simeq {}^{\mathbb{Z}[\alpha]}\!/\!_{\mathfrak{p} \cap \mathbb{Z}[\alpha]}$ hence $N(\mathfrak{p}_i) = p^{\deg \overline{\varphi}_i} = p^{f_i}$ for $\mathfrak{p}_i$ as above.

We now prove the claim. The map ${}^{\mathcal{O}_L}\!/\!_{\mathbb{Z}[\alpha]} \to {}^{\mathcal{O}_L}\!/\!_{\mathbb{Z}[\alpha]}$ given by multiplication by p is an isomorphism. It is injective as the kernel is a p-group so must be trivial, and ${}^{\mathcal{O}_L}\!/\!_{\mathbb{Z}[\alpha]}$ is a finite abelian group, so this is an isomorphism. But the kernel of the map $\iota \colon {}^{\mathbb{Z}[\alpha]}\!/\!_{p\mathbb{Z}[\alpha]} \to {}^{\mathcal{O}_L}\!/\!_{p\mathcal{O}_L}$ is ${}^{\mathbb{Z}[\alpha] \cap p\mathcal{O}_L}\!/\!_{p\mathbb{Z}[\alpha]}$, which is precisely the kernel of the map given by multiplication by p. So ι is injective.

ι is surjective if $\mathcal{O}_L = \mathbb{Z}[\alpha] + p\mathcal{O}_L$. The map given by multiplication by p is surjective, so ι is indeed surjective, and hence an isomorphism as required.

We have now constructed prime ideals $\mathfrak{p}_i = (p, \varphi_i(\alpha)) \lhd \mathcal{O}_L$ containing p with norm $N(\mathfrak{p}_i) = p^{f_i}$ with $f_i = \deg \overline{\varphi}_i$. We must now show that there are no other ideals containing p. Now, $\mathfrak{p}_i^{e_i} = (p, \varphi_i(\alpha))^{e_i} \subseteq (p, \varphi_i(\alpha)^{e_i})$, so

$$\mathfrak{p}_1^{e_1} \dots \mathfrak{p}_r^{e_r} \subseteq (p, \varphi_1(\alpha)^{e_1}) \dots (p, \varphi_r(\alpha)^{e_r}) \subseteq (p, \varphi_1(\alpha)^{e_1} \dots \varphi_r(\alpha)^{e_r})$$

But $\overline{\varphi_1^{e_1} \dots \varphi_r^{e_r}} = \overline{g}$, so $\varphi_1^{e_1} \dots \varphi_r^{e_r} = g + ps$. So $(p, \varphi_1(\alpha)^{e_1} \dots \varphi_r(\alpha)^{e_r}) = (p, g(\alpha)) = (p)$ as $g(\alpha) = 0$. So $\mathfrak{p}_1^{e_1} \dots \mathfrak{p}_r^{e_r} \subseteq (p)$. But $[L : \mathbb{Q}] = n = \deg g = \deg \overline{g} = \sum_{i=1}^{r} e_i \deg \overline{\varphi}_i = \sum_{i=1}^{r} e_i f_i$.

Taking norms,

$$N(\mathfrak{p}_1^{e_1} \dots \mathfrak{p}_r^{e_r}) = \prod_{i=1}^{r} N(\mathfrak{p}_i)^{e_i} = p^{e_1 f_1 + \dots + e_r f_r} = p^n = N((p))$$

One can show that if $\mathfrak{a} \subseteq \mathfrak{b}$ and $N(\mathfrak{a}) = N(\mathfrak{b})$, then $\mathfrak{a} = \mathfrak{b}$. So the two ideals are equal.

Note that if $i \neq j$, $\overline{\varphi}_i, \overline{\varphi}_j$ are coprime in $\mathbb{F}_p[x]$, so $\mathfrak{p}_i + \mathfrak{p}_j = (p, \varphi_i(\alpha), \varphi_j(\alpha)) \neq \mathfrak{p}_i$, so $\mathfrak{p}_i \neq \mathfrak{p}_j$. □

Note that since we choose a monic polynomial, $\deg \varphi_i(x) = \deg \overline{\varphi}_i(x)$. Different choices of $\varphi_i(x)$ give the same ideal as p is in the ideal. $\mathfrak{p}_i \neq \mathfrak{p}_j$ if $i \neq j$, and $\mathfrak{p}_1^{e_1} \dots \mathfrak{p}_r^{e_r}$ is the factorisation of (p) into irreducibles.

Remark. Most $\alpha \in \mathcal{O}_L$ have ${}^{\mathcal{O}_L}\!/\!{}_{\mathbb{Z}[\alpha]}$ finite, but the condition that $p \nmid \left| {}^{\mathcal{O}_L}\!/\!{}_{\mathbb{Z}[\alpha]} \right|$ is restrictive.

Example. Let $L = \mathbb{Q}(\sqrt{-11})$, and let us factorise $(5) \subseteq \mathcal{O}_L$. As $-11 \equiv 1 \bmod 4$, $\mathbb{Z}[\sqrt{-11}] \neq \mathcal{O}_L$. So $\mathbb{Z}[\sqrt{-11}]$ has index 2 in $\mathcal{O}_L$, and $5 \nmid 2$, so Dedekind's theorem applies. Modulo 5, $x^2 + 1 = (x-2)(x+2)$, so $(5) = (5, -2 + \sqrt{-11})(5, -2 - \sqrt{-11})$.

Example. In general, let $L = \mathbb{Q}(\sqrt{d})$ where d is square free and not equal to zero or one. Let p be an odd prime. Then, $\mathbb{Z}[\sqrt{d}] \subseteq \mathcal{O}_L$ has index 1 or 2, and both are coprime to p. Factorising $x^2 - d$ modulo p, there are three cases.

- Suppose there are two distinct roots modulo p of $x^2 - d$. Then, using the Legendre symbol, $\left(\frac{d}{p}\right) = 1$. In this case, $x^2 - d = (x-r)(x+r)$ for some $r \in \mathbb{Z}$. By Dedekind's theorem, $p = \mathfrak{p}_1\mathfrak{p}_2$ where $\mathfrak{p}_1 = (p, \sqrt{d} - r)$ and $\mathfrak{p}_2 = (p, \sqrt{d} + r)$. In this case, $N(\mathfrak{p}_1) = N(\mathfrak{p}_2) = p$; we say p *splits* in $L/\mathbb{Q}$.
- Suppose $x^2 - d$ is irreducible modulo p. Then $\left(\frac{d}{p}\right) = -1$. $(p) = \mathfrak{p}$ is prime; we say p is *inert* in L.
- Suppose $x^2 - d$ has a repeated root, so $d \equiv 0$ modulo p. Then $\left(\frac{d}{p}\right) = 0$. In this case, Dedekind's theorem gives $(p) = \mathfrak{p}^2$ where $\mathfrak{p} = (p, \sqrt{d})$. We say that p *ramifies* in L.

Now consider the case $p = 2$.

Lemma. 2 splits in L if and only if $d \equiv 1 \bmod 8$. 2 is inert in L if and only if $d \equiv 5 \bmod 8$. 2 ramifies in L if and only if $d \equiv 2, 3 \bmod 4$.

Proof. If $d \equiv 1 \bmod 4$, then $\mathcal{O}_L = \mathbb{Z}[\alpha]$ where $\alpha = \frac{1}{2}(1 + \sqrt{d})$. The minimal polynomial of α is $x^2 - x + \frac{1}{4}(1-d)$. Reducing modulo 2, if $d \equiv 1 \bmod 8$ then this is $x(x+1)$ so 2 splits. If $d \equiv 5 \bmod 8$ then this gives $x^2 + x + 1$ which is irreducible, so 2 is inert. If $d \equiv 2, 3 \bmod 4$, then $\mathcal{O}_L = \mathbb{Z}[\sqrt{d}]$ and $x^2 - d$ is either x^2 or $(x-1)^2$, which ramifies. □

Recall that $D_L = 4d$ if $d \equiv 2, 3 \bmod 4$, and $D_L = d$ if $d \equiv 1 \bmod 4$.

Corollary. Let $L = \mathbb{Q}(\sqrt{d})$. $p \mid D_L$ if and only if p ramifies in L.

Proof. Case analysis. □

Definition. Let $(p) = \mathfrak{p}_1^{e_1} \dots \mathfrak{p}_r^{e_r}$ be the factorisation of (p) into irreducibles in $\mathcal{O}_L$, where $p^{f_i} = N(\mathfrak{p}_i)$. We say that

- *p ramifies* if some e_i is greater than 1;
- *p is inert* if $r = 1$ and $e_1 = 1$, so (p) remains prime;
- *p splits* or *splits completely* if $r = n$ and $e_i = f_i = \dots = e_n = f_n = 1$.

Corollary. Let p be a prime and $p < n = [L : \mathbb{Q}]$. Let $\mathbb{Z}[\alpha] \subseteq \mathcal{O}_L$ have finite index coprime to p. Then p does not split completely.

Proof. Let g be the minimal polynomial of α. Suppose p splits, so g has n distinct roots in $\mathbb{F}_p$ by Dedekind's theorem. But $n > p$, so this is not possible. □

Example. Let $L = \mathbb{Q}(\alpha)$ and α has minimal polynomial $x^3 - x^2 - 2x - 8$. On an example sheet, we show that 2 splits completely in $\mathcal{O}_L$. Hence, for all $\beta \in \mathcal{O}_L \setminus \mathbb{Z}$, $\mathbb{Z}[\beta] \subseteq \mathcal{O}_L$ has even index.

Note that Dedekind's theorem allows for the factorisation of (p) for all but finitely many p, as if $\alpha \in \mathcal{O}_L$ with ${}^{\mathcal{O}_L}\!/_{\mathbb{Z}[\alpha]}$ finite, only finitely many primes p divide its order.

Theorem. For all primes p, we have $(p) = \mathfrak{p}_1^{e_1} \dots \mathfrak{p}_r^{e_r}$ with ${}^{\mathcal{O}_L}\!/_{\mathfrak{p}_i} = {}^{\mathbb{F}_p[x]}\!/_{\overline{\varphi}_i(x)}$ where $\overline{\varphi}_i \in \mathbb{F}_p[x]$ is an irreducible polynomial of degree f_i and $N(\mathfrak{p}_i) = p^{f_i}$, and ${}^{\mathcal{O}_L}\!/_{p\mathcal{O}_L} \simeq \prod_{i=1}^r {}^{\mathbb{F}_p[x]}\!/_{\overline{\varphi}_i(x)} = \prod_{i=1}^r \mathbb{F}_{p^{f_i}}$.

Dedekind's theorem implies that this holds if there exists $\alpha \in \mathcal{O}_L$ with $p \nmid \left|{}^{\mathcal{O}_L}\!/_{\mathbb{Z}[\alpha]}\right| < \infty$.

Theorem. p ramifies in L if and only if $p \mid D_L$.

3. Geometry of numbers

3.1. Imaginary quadratic fields

Let $L = \mathbb{Q}(\sqrt{d})$ where d is square-free and $d < 0$. $\mathcal{O}_L = \mathbb{Z}[\alpha]$ where $\alpha = \frac{1}{2}(1+\sqrt{d})$ or $\alpha = \sqrt{d}$. Choose a square root of d in $\mathbb{C}$ to construct an embedding of $\mathcal{O}_L$ into $\mathbb{C}$.

Suppose $\Lambda = \mathbb{Z}v_1 + \mathbb{Z}v_2 \subseteq \mathbb{R}^2$ where $\mathbb{R}^2$ is equipped with the Euclidean norm, and v_1, v_2 are linearly independent over $\mathbb{R}$. Let $A(\Lambda)$ be the area of the parallelogram generated by v_1 and v_2. If $v_i = a_i e_1 + b_i e_2$, we have

$$A(\Lambda) = \left|\det\begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix}\right|$$

Minkowski's lemma is that a closed disk S around zero contains a nonzero point of Λ whenever the area of S is at least $4A(\Lambda)$. More precisely, there exists $\alpha \in \Lambda$ such that $0 < |\alpha|^2 < \frac{4A(\Lambda)}{\pi}$. Note that this condition depends only on the area of the parallelogram, not its shape. This will be proven shortly.

We will apply this to $\Lambda = \mathfrak{a} \trianglelefteq \mathcal{O}_L$ for $L = \mathbb{Q}(\sqrt{d})$, $d < 0$ square-free. Let $\sqrt{d} \in \mathbb{C}$ be chosen with positive imaginary part to embed $\mathcal{O}_L$ in $\mathbb{C}$.

Lemma. (i) if $\alpha = a + b\sqrt{d} \in \mathcal{O}_L$, then $|\alpha|^2 = (a+b\sqrt{d})(a-b\sqrt{d}) = N(\alpha)$;

(ii) $A(\mathcal{O}_L) = \frac{1}{2}\sqrt{|D_L|}$;

(iii) $A(\mathfrak{a}) = N(\mathfrak{a})A(\mathcal{O}_L)$;

(iv) $A(\mathfrak{a}) = \frac{1}{2}|\Delta(\alpha_1,\alpha_2)|^{\frac{1}{2}}$ where α_1, α_2 are an integral basis for $\mathfrak{a}$.

Proof. Part (i) is clear. (iv) implies (ii) and (iii). We will prove (iv) later in a more general setting, giving the justification for the coefficient $\frac{1}{2}$.

We now prove (ii) and (iii) manually, without appealing to (iv). For part (ii), $\mathcal{O}_L$ has basis $1, \alpha$. Therefore, $A(\mathcal{O}_L) = \frac{1}{2}\sqrt{d}$ or $\sqrt{d}$, which is exactly $\frac{1}{2}\sqrt{|D_L|}$. Part (iii) is a variant of the fact that $\Delta(\alpha_1,\dots,\alpha_n) = N(\mathfrak{a})^2 D_L$. □

Minkowski's lemma implies that there exists $\alpha \in \mathfrak{a}$ with $N(\alpha) \leq \frac{4A(\mathfrak{a})}{\pi} = N(\mathfrak{a})C_L$ where $C_L = \frac{2\sqrt{|D_L|}}{\pi}$ is Minkowski's constant. Since $\alpha \in \mathfrak{a}$, $(\alpha) \subseteq \mathfrak{a}$. Hence $(\alpha) = \mathfrak{a}\mathfrak{b}$ for some $\mathfrak{b} \trianglelefteq \mathcal{O}_L$. So $N(\alpha) = N((\alpha)) = N(\mathfrak{a})N(\mathfrak{b})$, so $N(\mathfrak{b}) \leq C_L$.

Recall that the class group of L is ${}^{I_L}\!/\!_{P_L}$, the quotient of fractional ideals over principal ideals. Then, $[\mathfrak{b}] = [\mathfrak{a}^{-1}] \in \mathrm{Cl}_L$. Replacing $\mathfrak{a}$ with $\mathfrak{a}^{-1}$, we have shown that for all $[\mathfrak{a}] \in \mathrm{Cl}_L$, there exists a representative $\mathfrak{b}$ of $[\mathfrak{a}]$ which is an ideal with $N(\mathfrak{b}) \leq \frac{2\sqrt{|D_L|}}{\pi} = C_L$. But for all $m \in \mathbb{Z}$, the number of ideals $\mathfrak{a} \trianglelefteq \mathcal{O}_L$ with $N(\mathfrak{a}) = m$ is finite; indeed, if $N(\mathfrak{a}) = m$, then $m \in \mathfrak{a}$

so $\mathfrak{a} \mid (m)$, but there are only finitely many ideals dividing (m), as they biject with ideals in ${}^{\mathcal{O}_L}\!/\!{}_{m\mathcal{O}_L} \simeq \left({}^{\mathbb{Z}}\!/\!{}_{m\mathbb{Z}}\right)^n$.

Therefore, we have shown that Cl_L is finite, and generated by the class of prime ideals dividing (p), for p a prime integer less than $\frac{2\sqrt{|D_L|}}{\pi} = C_L$. Indeed, if $\mathfrak{a} = \mathfrak{p}_1^{e_1} \dots \mathfrak{p}_r^{e_r}$ with $N(\mathfrak{a}) < C_L$, then $N(\mathfrak{p}_i) < C_L$.

Example. Let $d = -7$. Then $D_L = -7$, and $\frac{2\sqrt{7}}{\pi} < 2$. So there are no primes $p < C_L$, giving $\mathrm{Cl}_L = \{1\}$. In particular, $\mathcal{O}_L$ is a unique factorisation domain. Similarly, $d = -1, -2, -3$ give unique factorisation domains.

Example. Let $d = -5$. Here, $D_L = -20$, and $2 < \frac{4\sqrt{5}}{\pi} < 3$. Hence, Cl_L is generated by prime ideals dividing (2). Note that $(2) = (2, 1 + \sqrt{-5})^2$ by Dedekind's theorem.

We now must check if $(2, 1 + \sqrt{-5})$ is principal. If $(2, 1 + \sqrt{-5}) = (\beta)$, then $N(\beta) = 2$. But $\beta = a + b\sqrt{-5}$, so $N(\beta) = a^2 + 5b^2$, which is not satisfiable by integers. So $(2, 1 + \sqrt{-5})$ is principal but its square is, so $\mathrm{Cl}_L = {}^{\mathbb{Z}}\!/\!{}_{2\mathbb{Z}}$.

Example. Let $d = -17$, then $5 < C_L < 6$. Cl_L is generated by prime ideals dividing $(2), (3), (5)$. Modulo 2, $x^2 + 17 = x^2 + 1 = (x + 1)^2$, so $(2) = \mathfrak{p}^2$ where $\mathfrak{p} = (2, 1 + \sqrt{-17})$. Modulo 3, $x^2 + 17 = x^2 - 1 = (x + 1)(x - 1)$, giving $(3) = \mathfrak{q}\overline{\mathfrak{q}}$ where $\mathfrak{q} = (3, 1 + \sqrt{-17}), \overline{\mathfrak{q}} = (3, 1 - \sqrt{-17})$. Modulo 5, $x^2 + 17 = x^2 + 2$ which is irreducible, so (5) is inert, so is trivial in the class group.

Hence $\mathrm{Cl}_L = (\mathfrak{p}, \mathfrak{q}, \overline{\mathfrak{q}}) = (\mathfrak{p}, \mathfrak{q})$. We could compute powers of $\mathfrak{p}$ and $\mathfrak{q}$ until we obtain all nontrivial relations between them. A more efficient way to compute Cl_L in this case is to find principal ideals of small norm which are multiples of 2 and 3 to find the relations. Consider $(1 + \sqrt{-17})$, which has norm $N(1 + \sqrt{-17}) = 18 = 2 \cdot 3^2$. Note that $1 + \sqrt{-17} \in \mathfrak{p} \cap \mathfrak{q}$ so $(1 + \sqrt{-17}) = \mathfrak{p}\mathfrak{q}\mathfrak{r}$ where $\mathfrak{r} \in (\mathfrak{p}, \mathfrak{q})$. We can show that $\mathfrak{r} = \mathfrak{q}$. This shows that $[\mathfrak{p}] = [\mathfrak{q}]^{-2}$ in Cl_L. So Cl_L is generated by $[\mathfrak{q}]$. So it is cyclic, and we can show $[\mathfrak{q}]^2 \neq 1$, as $\mathfrak{p}$ is not principal, but $[\mathfrak{q}]^4 = [\mathfrak{p}^2]^{-1} = 1$. So $\mathrm{Cl}_L = {}^{\mathbb{Z}}\!/\!{}_{4\mathbb{Z}}$.

Theorem. Let $L = \mathbb{Q}(\sqrt{-d})$ with $d > 0$.

(i) $\mathcal{O}_L$ is a unique factorisation domain if $d \in \{1, 2, 3, 7, 11, 19, 43, 67, 163\}$;

(ii) there are no others.

3.2. Lattices

Definition. A subset $X \subseteq \mathbb{R}^n$ is called *discrete* if for all $K \subseteq \mathbb{R}^n$ compact, $K \cap X$ is finite. Equivalently, for all $x \in X$ there exists $\varepsilon > 0$ with $B_\varepsilon(x) \cap X = \{x\}$.

Recall that $K \subseteq \mathbb{R}^n$ is compact if and only if it is closed and bounded.

Proposition. Let $\Lambda \subseteq \mathbb{R}^n$. Then the following are equivalent.

(i) Λ is a discrete subgroup of $(\mathbb{R}^n, +)$;

(ii) $\Lambda = \left\{\sum_{i=1}^m n_i x_i \mid n_i \in \mathbb{Z}\right\}$ where $x_1, \dots, x_m$ are linearly independent over $\mathbb{R}$.

Example. $\mathbb{Z}\sqrt{2} + \mathbb{Z}\sqrt{3} \subseteq \mathbb{R}$ is not discrete. If $\Lambda = \mathfrak{a} \trianglelefteq O_L$ is an ideal where $L = \mathbb{Q}(\sqrt{-d})$ and $d > 0$, Λ is discrete.

Proof. (ii) implies (i). Observe that if $g \in GL_n(\mathbb{R})$, then $g\Lambda$ is discrete if Λ is. $g\Lambda$ satisfies (ii) if and only if Λ does. Suppose property (ii) holds, so $\Lambda = \left\{\sum_{i=1}^m n_i x_i \mid n_i \in \mathbb{Z}\right\}$. There exists $g \in GL_n(\mathbb{R})$ such that $gx_i = e_i$ where the e_i form the standard basis of $\mathbb{R}^n$. Clearly, $\bigoplus_{i=1}^m \mathbb{Z}e_i$ is discrete.

(i) implies (ii). Let $y_1, \dots, y_m \in \Lambda$ which are $\mathbb{R}$-linearly independent such that m is maximal. Note that $m \leq n$. Also,

$$\left\{\sum_{i=1}^m \lambda_i y_i \,\middle|\, \lambda_i \in \mathbb{R}\right\} = \left\{\sum_{i=1}^N \lambda_\alpha z_\alpha \,\middle|\, \lambda_\alpha \in \mathbb{R}, z_\alpha \in \Lambda, N \geq 0\right\}$$

This is the smallest $\mathbb{R}$-vector subspace of $\mathbb{R}^n$ containing Λ. Let $X = \left\{\sum_{i=1}^m \lambda_i y_i \mid \lambda_i \in [0,1]\right\}$. This is closed and bounded, hence compact. Λ is discrete, so $X \cap \Lambda$ is finite.

Consider the subgroup $\mathbb{Z}^m = \bigoplus_{i=1}^m \mathbb{Z}y_i \subseteq \Lambda$. We can write $\lambda \in \Lambda$ as $\lambda = \lambda_0 + \lambda_1$ where $\lambda_0 \in X \cap \Lambda$ is the integral part and $\lambda_1 \in \mathbb{Z}^m = \bigoplus_{i=1}^m \mathbb{Z}y_i$ is the fractional part. Hence, $\left|{}^{\Lambda}\!/_{\mathbb{Z}^m}\right| \leq |X \cap \Lambda|$ is finite. Let $d = \left|{}^{\Lambda}\!/_{\mathbb{Z}^m}\right|$, so by Lagrange's theorem, $d = 0$ in ${}^{\Lambda}\!/_{\mathbb{Z}^m}$, so $d\Lambda \subseteq \mathbb{Z}^m$. In particular, $\mathbb{Z}^m \subseteq \Lambda \subseteq \frac{1}{d}\mathbb{Z}^m$. The structure theorem for finitely generated abelian groups shows that there exist $x_1, \dots, x_m \in \Lambda$ with $\Lambda = \bigoplus_{i=1}^m \mathbb{Z}x_i$. □

Definition. If rank $\Lambda = n$, so if $n = m$, we say Λ is a *lattice* in $\mathbb{R}^n$.

Definition. Let $\Lambda \subseteq \mathbb{R}^n$ be a lattice with basis $x_1, \dots, x_n$. The *fundamental parallelogram* is $P = \left\{\sum_{i=1}^n \lambda_i x_i \mid \lambda_i \in [0,1]\right\}$. The *covolume* of Λ is the volume of P, which is $|\det A|$ if $x_i = \sum_{j=1}^n a_{ij} e_j$.

Note that if $x_1', \dots, x_n'$ are another basis of Λ, the change of basis matrix B given by $x_i' = \sum_{j=1}^m b_{ij} x_j$ has integer coefficients, so $B \in GL_n(\mathbb{Z})$, giving $\det B = \pm 1$. Hence, the covolume is well-defined irrespective of the choice of basis. Observe that P is a fundamental domain for the action of Λ on $\mathbb{R}^n$; $\mathbb{R}^n = \bigcup_{\gamma \in \Lambda} (\gamma + P)$ and $(\gamma + P) \cap (\mu + P) \subseteq \partial P$ if $\gamma \neq \mu$. We can think of P as a set of coset representatives for ${}^{\mathbb{R}^n}\!/_{\Lambda}$, ignoring the boundary of P; this can be justified by noting that ∂P has no volume.

3.3. Minkowski's lemma

Theorem. Let $\Lambda \subseteq \mathbb{R}^n$ be a lattice, and P be a fundamental parallelogram for it. Let $S \subseteq \mathbb{R}^n$ be a measurable set.

(i) If $\mathrm{vol}(S) > \mathrm{covol}(\Lambda)$, there exist $x, y \in S$ with $x \neq y$ and $x - y \in \Lambda$.

(ii) Suppose $s \in S$ if and only if $-s \in S$, so S is *symmetric around zero*, and that S is convex. Then, if

(a) $\mathrm{vol}(S) > 2^n \,\mathrm{covol}(\Lambda)$, or

(b) $\mathrm{vol}(S) \geq 2^n \,\mathrm{covol}(\Lambda)$ and S is closed,

then there exists $\gamma \in S \cap \Lambda$ with $\gamma \neq 0$.

Note that this implies the result we used when $n = 2$. In the case of the square lattice $\Lambda = \mathbb{Z}^n$ and $S = [-1, 1]^n$, we can see that these bounds are sharp.

Proof. Part (i). Observe that $\mathrm{vol}(S) = \sum_{\gamma\in\Lambda} \mathrm{vol}(S \cap (P + \gamma))$ as P is a fundamental domain, volume is additive, and $\mathrm{vol}(\partial(P+\gamma)) = 0$. Note that $\mathrm{vol}(S\cap(P+\gamma)) = \mathrm{vol}((S-\gamma)\cap P)$ as volume is translation invariant. We claim that the sets $(S - \gamma) \cap P$ are not pairwise disjoint. Indeed, if they were, then $\mathrm{vol}(P) \geq \sum_{\gamma\in\Lambda} \mathrm{vol}((S - \gamma) \cap P) = \mathrm{vol}(S)$ contradicting the assumption. Hence there exists $\gamma\mu \in \Lambda$ with $\gamma \neq \mu$ such that $(S - \gamma) \cap P$ and $(S - \mu) \cap P$ are not disjoint, so there exist $x, y \in S$ with $x - \gamma = y - \mu$, hence $x - y \in \Lambda$.

Part (ii)(a). Let $S' = \frac{1}{2}S = \left\{\frac{1}{2}s \mid s \in S\right\}$. Then $\mathrm{vol}(S') = 2^{-n}\,\mathrm{vol}(S) > \mathrm{covol}(\Lambda)$ by assumption. By part (i), there exist $y, z \in S'$ with $y - z \in \Lambda \setminus \{0\}$. But $y - z = \frac{1}{2}(2y + -2z)$. $2z \in S$ so $-2z \in S$ as S is symmetric around zero. $2y \in S$, and S is convex, so $y - z \in S$ as required.

Part (ii)(b). Apply part (ii)(a) to $S_m = \left(1 + \frac{1}{m}\right)S$ for all $m \in \mathbb{N}, m > 0$. We obtain $\gamma_m \in S_m \cap \Lambda$ with $\gamma_m \neq 0$. By convexity of S, $S_m \subseteq S_1$. So $\gamma_1, \gamma_2, \dots$ are contained in $S_1 \cap \Lambda$, which is a finite set as S_1 is closed and bounded (without loss of generality) and Λ is discrete. So there exists $\gamma \in S_m \cap \Lambda$ such that $\gamma_m = \gamma$ for infinitely many m. Hence, $\gamma \in \bigcap_{m>0} S_m = S$ as S is closed. Therefore $\gamma \in S \cap \Lambda$ with $\gamma \neq 0$. □

Let L be a number field and let $n = [L : \mathbb{Q}]$. Let $\sigma_1, \dots, \sigma_r : L \to \mathbb{R}$ be the real embeddings, and $\sigma_{r+1}, \dots, \sigma_{r+s}, \overline{\sigma_{r+1}}, \dots, \overline{\sigma_{r+s}} : L \to \mathbb{C}$ be the complex embeddings, where $r + 2s = n$. This gives an embedding

$$(\sigma_1, \dots, \sigma_{r+s}) \colon L \hookrightarrow \mathbb{R}^r \times \mathbb{C}^s \xrightarrow{\simeq} \mathbb{R}^r \times \mathbb{R}^{2s} = \mathbb{R}^{r+2s}$$

In other words, we can write

$$\sigma = (\sigma_1, \dots, \sigma_r, \mathrm{Re}\,\sigma_{r+1}, \mathrm{Im}\,\sigma_{r+1}, \dots, \mathrm{Re}\,\sigma_{r+s}, \mathrm{Im}\,\sigma_{r+s})$$

Lemma. $\sigma(\mathcal{O}_L)$ is a lattice in $\mathbb{R}^n$ of covolume $2^{-s}|D_L|^{\frac{1}{2}}$. If $\mathfrak{a} \trianglelefteq \mathcal{O}_L$ is an ideal, then $\sigma(\mathfrak{a})$ is a lattice, and $\mathrm{covol}(\sigma(\mathfrak{a})) = 2^{-s}|D_L|^{\frac{1}{2}} N(\mathfrak{a})$.

Proof. The first part is a special case of the second part. Recall that $\mathfrak{a}$ has an integral basis $\gamma_1, \dots, \gamma_n$, and $(\det(\sigma_i(\gamma_j)))^2 = \Delta(\gamma_1, \dots, \gamma_n) = N(\mathfrak{a})^2 D_L$. Hence, $\left|\det(\sigma_i(\gamma_j))\right| = N(\mathfrak{a})|D_L|^{\frac{1}{2}}$. Note that if $\sigma_{r+i}(\gamma)\overline{\sigma_{r+i}(\gamma)} = z\overline{z}$,

$$\begin{pmatrix} \operatorname{Re} z \\ \operatorname{Im} z \end{pmatrix} = \begin{pmatrix} \frac{1}{2}(z+\overline{z}) \\ \frac{1}{2i}(z-\overline{z}) \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix}\begin{pmatrix} z \\ \overline{z} \end{pmatrix}$$

The determinant of the change of basis matrix is $-\frac{1}{2}$. □

Proposition (Minkowski bound). Let $\mathfrak{a} \trianglelefteq \mathcal{O}_L$. Then there exists $\alpha \in \mathfrak{a}$ with $\alpha \neq 0$ and $|N(\alpha)| \leq C_L N(\mathfrak{a})$ where $C_L = \left(\frac{4}{\pi}\right)^s \frac{n!}{n^n}|D_L|^{\frac{1}{2}}$.

Proof. Let

$$B_{r,s}(t) = \left\{(y_1, \dots, y_r, z_1, \dots, z_s) \in \mathbb{R}^r \times \mathbb{C}^s \mid \sum |y_i| + 2|z_i| \leq t\right\}$$

This set is closed and bounded, hence compact. It is also convex, symmetric around zero, and measurable with volume $2^r\left(\frac{\pi}{2}\right)^2 \frac{t^n}{n!}$. Choose t such that the volume of $B_{r,s}(t)$ is $2^n \operatorname{covol}(\mathfrak{a})$, so $t^n = \left(\frac{4}{\pi}\right)^s n!|D_L|^{\frac{1}{2}}N(\mathfrak{a})$. Minkowski's lemma implies that there exists $\alpha \in \mathfrak{a}$ and $\alpha \neq 0$ such that $\sigma(\alpha) = (y_1, \dots, y_r, z_1, \dots, z_s) \in B_{r,s}(t)$.

Note that $N(\alpha) = y_1 \dots y_r z_1 \overline{z_1} \dots z_s \overline{z_s} = \prod y_i \prod \left|z_j\right|^2$. Since the geometric mean is at most the arithmetic mean, taking nth roots we obtain $|N(\alpha)|^{\frac{1}{n}} \leq \frac{1}{n}\left(\sum |y_i| + 2\sum \left|z_j\right|\right) \leq \frac{t}{n}$ as $\sigma(\alpha) \in B_{r,s}(t)$. So $|N(\alpha)| \leq \frac{t^n}{n^n} = C_L N(\mathfrak{a})$ as required. □

To show that the volume of $B_{r,s}(t)$ is $2^r\left(\frac{\pi}{2}\right)^2 \frac{t^n}{n!}$, we can use induction with base cases $B_{1,0}(t) = [-t, t]$ and $B_{0,1}(t) = \frac{\pi}{4}t^2$. Given the result for $B_{r,s}(t)$, the volume of $B_{r+1,s}(t)$ is

$$\int_{-t}^{t} \operatorname{vol} B_{r,s}(t - |y|)\,\mathrm{d}y = 2\int_0^t \left(\frac{\pi}{2}\right)^s 2^r \frac{(ty)^n}{n!}\,\mathrm{d}y = 2^{r+1}\left(\frac{\pi}{2}\right)^2 \frac{t^{n+1}}{n!}$$

The other inductive step is on an example sheet.

Corollary. Every element of the class group $[\mathfrak{a}]$ has a representative $\mathfrak{a} \trianglelefteq \mathcal{O}_L$ with norm at most C_L.

Theorem. The class group of L is finite, and generated by prime ideals $\mathfrak{a} \trianglelefteq \mathcal{O}_L$ with $N(\mathfrak{a}) \leq C_L$.

Proof. Follows the argument used for imaginary quadratic fields. □

Theorem (Hermite, Minkowski)**.** Let $n \geq 2$. Then $|D_L| \geq \frac{\pi}{3}\left(\frac{3\pi}{4}\right)^{n-1} > 1$. In particular, $|D_L| > 1$, so at least one prime ramifies in L.

Proof. Apply this to $[\mathcal{O}_L]$ and obtain an ideal $\mathfrak{a} \trianglelefteq \mathcal{O}_L$ with $1 \leq N(\mathfrak{a}) \leq C_L$, so $C_L \geq 1$. So

$$|D_L|^{\frac{1}{2}} \geq \left(\frac{\pi}{4}\right)^s \frac{n^n}{n!} \geq \left(\frac{\pi}{4}\right)^{\frac{n}{2}} \frac{n^n}{n!} a_n^{\frac{1}{2}}$$

as $\frac{\pi}{4} < 1$ and $s \leq \frac{n}{2}$. So $a_2 = \frac{\pi^2}{4}$ and $\frac{a_{n+1}}{a_n} = \frac{\pi}{4}\left(1 + \frac{1}{n}\right)^{2n} > \frac{\pi}{4}(1+2) = \frac{3\pi}{4}$. So $a_n \geq \frac{\pi^2}{4}\left(\frac{3\pi}{4}\right)^{n-2} = \frac{\pi}{3}\left(\frac{3\pi}{4}\right)^{n-1}$. □

4. Dirichlet's unit theorem

4.1. Real quadratic fields

Recall that $\alpha \in \mathcal{O}_L$ is a unit if and only if $N(\alpha) = \pm 1$. We aim to show that $\mathcal{O}_L^\star \simeq \boldsymbol{\mu}_L \times \mathbb{Z}^{r+s-1}$ where $\boldsymbol{\mu}_L = \{\alpha \in L \mid \alpha^a = 1 \text{ for some } a > 0\}$ is the set of roots of unity in L, a finite cyclic group.

Example. Let $L = \mathbb{Q}(\sqrt{d})$ where $d > 0$ is square-free. Here, $r = 2, s = 0, n = 2$. $L \subseteq \mathbb{R}$ gives $\boldsymbol{\mu}_L \subseteq \{\pm 1\}$ so $\boldsymbol{\mu}_L = \{\pm 1\}$. Note that $N(x + y\sqrt{d}) = x^2 - dy^2$, so Dirichlet's theorem implies the following statement, which we will now prove directly.

Theorem (Pell's equation)**.** There exist infinitely many $x + y\sqrt{d} \in \mathcal{O}_L$ with $x^2 - dy^2 = \pm 1$.

Proof. Recall that we have $\sigma \colon \mathcal{O}_L \to \mathbb{R}^2$ given by $x + y\sqrt{d} \mapsto (x + y\sqrt{d}, x - y\sqrt{d})$. For example, if $d = 2$, the image is a lattice with basis $(1,1), (-\sqrt{2}, \sqrt{2})$, note also that no point lies in the coordinate axes apart from 0. The covolume of $\sigma(\mathcal{O}_L)$ is $|D_L|^{\frac{1}{2}}$.

Consider

$$S_t = \left\{ (y_1, y_2) \in \mathbb{R}^2 \;\middle|\; |y_1| \leq t, |y_2| \leq \frac{|D_L|^{\frac{1}{2}}}{t} \right\}$$

The volume of S_t is $4|D_L|^{\frac{1}{2}} = 2^n \operatorname{covol}(\sigma(\mathcal{O}_L))$ as $n = 2$. Minkowski's lemma implies that there exists a nonzero $\alpha \in \mathcal{O}_L$ with $\sigma(\alpha) \in S_t$. But $\sigma(\alpha) = (y_1, y_2)$ gives $N(\alpha) = y_1 y_2$.

We have therefore found an element $\alpha \in \mathcal{O}_L$ with $\sigma(\alpha) \in S_t$ that has norm satisfying $1 \leq n(\alpha) \leq |D_L|^{\frac{1}{2}}$. We show that there exist infinitely many such α for $0 < t < 1$, so there are infinitely many $\alpha \in \mathcal{O}_L$ with $|N(\alpha)| = N((\alpha)) < |D_L|^{\frac{1}{2}}$. For fixed t, $S_t \cap \sigma(\mathcal{O}_L)$ is finite as S_t is compact. Given $t_1 > t_2 > \dots > t_n$, choose t_{n+1} less than all y_1 where $\sigma(\alpha) = (y_1, y_2) \in S_{t_n} \cap \sigma(\mathcal{O}_L)$. Note that $\alpha \neq 0$ so $\sigma_1(\alpha) \neq 0$, so $t_{n+1} > 0$.

Hence, there exists $m \in \mathbb{Z}$ with $1 \leq |m| \leq |D_L|^{\frac{1}{2}}$ for which there are infinitely many α with $N(\alpha) = m$, by the pigeonhole principle. But ideals $\mathfrak{a} \trianglelefteq \mathcal{O}_L$ with $m \in \mathfrak{a}$ biject with ideals in $\mathcal{O}_L/m = \left(\mathbb{Z}/m\mathbb{Z}\right)^2$, and hence there are finitely many of them. Again by the pigeonhole principle, there exists $\beta \in \mathcal{O}_L$ and infinitely many $\alpha \in \mathcal{O}_L$ with $N(\beta) = N(\alpha) = m$, where $(\beta) = (\alpha)$. But $\frac{\beta}{\alpha}$ is a unit, so there are infinitely many units. □

We can prove Dirichlet's unit theorem for real quadratic fields from this result.

Corollary. $\mathcal{O}_L^\star = \{\pm \varepsilon_0^n \mid n \in \mathbb{Z}\}$ for $\varepsilon_0 \in \mathcal{O}_L^\star$.

Such an ε_0 is called a *fundamental unit*.

Remark. As there are infinitely many units, there exists $\varepsilon \in \mathcal{O}_L^\star$ with $\varepsilon \neq \pm 1$. Hence, $|\sigma_1(\varepsilon)| \neq \pm 1$ as $\sigma_1(\varepsilon) = \pm 1$ if and only if $\varepsilon = \pm 1$. Replacing ε by ε^{-1} if necessary, we can assume

$E = |\sigma_1(\varepsilon)| > 1$. Consider $\{\alpha \in \mathcal{O}_L \mid N(\alpha) = \pm 1, 1 \leq |\sigma_1(\alpha)| \leq E\}$, which is a finite set as $\mathcal{O}_L$ is discrete in $\mathbb{R}^2$. Hence, ε_0 can be chosen in this set with minimum $|\sigma_1(\varepsilon_0)|$ and $\varepsilon_0 \neq \pm 1$.

We claim that if $\varepsilon \in \mathcal{O}_L^\star$ has $\sigma_1(\varepsilon) > 0$, then $\varepsilon = \varepsilon_0^N$ for some $N \in \mathbb{Z}$. Indeed, we can write $\frac{\log \sigma_1(\varepsilon)}{\log \sigma_1(\varepsilon_0)} = N + \gamma$ where $N \in \mathbb{Z}, 0 \leq \gamma < 1$. Hence $\varepsilon\varepsilon_0^{-N} = \varepsilon_0^\gamma$, and if $\gamma \neq 0$, $|\varepsilon_0^\gamma| = |\varepsilon|^\gamma < |\varepsilon_0|$ contradicting the choice of ε_0 (taking σ_1 as necessary to simplify notation).

4.2. General case

We can prove Dirichlet's unit theorem in general.

Let L be a number field and let $[L : \mathbb{Q}] = n$ with $\sigma_1, \dots, \sigma_r : L \to \mathbb{R}$ real embeddings and $\sigma_{r+1}, \dots, \sigma_{r+s}, \overline{\sigma}_{r+1}, \dots, \overline{\sigma}_{r+s} : L \to \mathbb{C}$ complex embeddings, choosing some representative between $\sigma_{r+i}, \overline{\sigma}_{r+i}$ arbitrarily. Define a map $\ell : \mathcal{O}_L^\star \to \mathbb{R}^{r+s}$ by

$$\ell(x) = (\log|\sigma_1(x)|, \dots, \log|\sigma_r(x)|, 2\log|\sigma_{r+1}(x)|, \dots, 2\log|\sigma_{r+s}(x)|)$$

Lemma. (i) The image of ℓ is a discrete subgroup of $\mathbb{R}^{r+s}$.

(ii) The kernel of ℓ is $\boldsymbol{\mu}_L$, the roots of unity in L, which is a finite cyclic group.

Remark. ℓ is independent of the choice of representative $\sigma_{r+i}, \overline{\sigma}_{r+i}$, as they have the same absolute value.

Proof. Part (i). $\log|ab| = \log|a| + \log|b|$, so ℓ is a group homomorphism. The image is therefore an additive subgroup of $\mathbb{R}^{r+s}$. For part (i), it suffices to show that $\operatorname{Im}\ell \cap [-A, A]^{r+s}$ is finite for all $A > 0$. ℓ factorises as

$$\begin{tikzcd} \mathcal{O}_L^\star \arrow[r, hook, "\sigma"] & (\mathbb{R}_{\neq 0})^r \times \mathbb{C}^s \arrow[r, "j"] & \mathbb{R}^{r+s} \end{tikzcd}$$

where

$$j(y_1, \dots, y_r, z_1, \dots, z_s) = (\log|y_1|, \dots, \log|y_r|, 2\log|z_1|, \dots, 2\log|z_s|)$$

and

$$j^{-1}([-A, A]^{r+s}) = \left\{(y_i, z_j) \mid e^{-A} \leq |y_i| \leq e^A, e^{-A} \leq 2|z_j| \leq e^A\right\}$$

which is compact. As $\sigma(\mathcal{O}_L)$ is a lattice, $\sigma(\mathcal{O}_L^\star) \cap j^{-1}([-A, A]^{r+s})$ is finite. This gives (i), and also shows that $\ker j = \ker \ell$ is finite.

Part (ii). $\ker \ell$ is a group and finite, so every element has finite order. In particular, $\ker \ell \leq \boldsymbol{\mu}_L$. But each root of unity lies in $\ker \ell$, so $\ker \ell = \boldsymbol{\mu}_L$. But $L \hookrightarrow \mathbb{C}$ by any embedding, so $\boldsymbol{\mu}_L$ is contained in the set of roots of unity in $\mathbb{C}$ of a fixed order, which is a cyclic group. Subgroups of cyclic groups are cyclic. □

Note that if $r > 0$, $L \hookrightarrow \mathbb{R}$, so $\boldsymbol{\mu}_L = \{\pm 1\}$.

Observe that $\operatorname{Im}\ell$ is contained in the set $\{(y_1,\dots,y_{r+s}) \mid y_1 + \dots + y_{r+s} = 0\}$. Indeed, $\alpha \in \mathcal{O}_L^\star$ gives $N(\alpha) = \prod_{i=1}^r \sigma_i(\alpha) \prod_{i=1}^s \sigma_{r+i}(\alpha)\overline{\sigma}_{r+i}(\alpha) = \pm 1$, so taking logarithms,

$$\log|N(\alpha)| = \sum_{i=1}^r \log|\sigma_i(\alpha)| + \sum_{i=1}^s 2\log|\sigma_{r+i}(\alpha)| = 0$$

So $\operatorname{Im}\ell \subseteq \mathbb{R}^{r+s-1}$ is a discrete subgroup, hence isomorphic to $\mathbb{Z}^a$ for $a \leq r+s-1$.

Theorem (Dirichlet's unit theorem)**.** $\operatorname{Im}\ell \subseteq \mathbb{R}^{r+s-1}$ is a lattice; it is isomorphic to $\mathbb{Z}^{r+s-1}$.

We now prove this theorem.

Lemma. Let $1 \leq k \leq s$, and $\alpha \in \mathcal{O}_L$, $\alpha \neq 0$. Then there exists $\beta \in \mathcal{O}_L$ with $|N(\beta)| \leq \left(\frac{2}{\pi}\right)^s |D_L|^{\frac{1}{2}}$ and with $b_i < a_i$ for all $i \neq k$, where $\ell(\alpha) = (a_1,\dots,a_{r+s})$ and $\ell(\beta) = (b_1,\dots,b_{r+s})$.

Proof. Apply Minkowski's lemma. Let

$$S = \left\{(y_1,\dots,y_r,z_r,\dots,z_s) \in \mathbb{R}^r \times \mathbb{C}^s \simeq \mathbb{R}^n \;\middle|\; |y_i| \leq c_i, |z_j|^2 \leq c_{r+j}\right\}$$

We have $\operatorname{vol}(S) = 2^r\pi^s c_1 \dots c_{r+s}$. This is convex and symmetric around zero. By choosing c_i such that $0 < c_i < e^{a_i}$ for $i \neq k$, and setting $c_k = \left(\frac{2}{\pi}\right)^s |D_L|^{\frac{1}{2}} c_1^{-1} \dots c_{k-1}^{-1} c_{k+1}^{-1} \dots c_{r+s}^{-1}$, Minkowski gives $\beta \in \sigma(\mathcal{O}_L) \cap S$. □

Fix some $1 \leq k \leq s$. Repeatedly applying this lemma, we can obtain a sequence $\alpha_1, \alpha_2, \dots \in \mathcal{O}_L$ such that $N(\alpha_j)$ is bounded, and for all $i \neq k$, the ith coordinate of $\ell(\alpha_1), \ell(\alpha_2), \dots$ is strictly decreasing. Hence, there exists $t < t'$ with $N(\alpha_t) = N(\alpha_{t'}) = m$ as there are only finitely many possible norms of the α_t, and $\alpha_t = \alpha_{t'}$ modulo ${}^{\mathcal{O}_L}\!/_{m}$ by the pigeonhole principle. Therefore $(\alpha_t) = (\alpha_{t'})$ as in the proof for real quadratic fields.

Let $u_k = \alpha_t\alpha_{t'}^{-1}$; this is a unit in $\mathcal{O}_L$ such that $\ell(u_k) = \ell(a_t) - \ell(a_{t'}) = (y_1,\dots,y_{r+s})$ has $y_i < 0$ if $i \neq k$. Note that as $\sum y_i = 0$, we have $y_k > 0$.

We now have units $u_1,\dots,u_{r+s}$ by performing this for each coordinate. We now show that $\ell(u_1),\dots,\ell(u_{r+s-1})$ are linearly independent, hence the rank of $\ell(\mathcal{O}_L^\star)$ is $r+s-1$. Indeed, let A be the $(r+s) \times (r+s)$ matrix with jth row given by $\ell(u_j)$, and apply the following lemma.

Lemma. Let $A \in M_{m\times m}(\mathbb{R})$ be a matrix with $a_{ii} > 0$, $a_{ij} < 0$ for $i \neq j$, and $\sum_j a_{ij} \geq 0$ for all i. Then $\operatorname{rank} A \geq m - 1$.

Note that the assumptions of this lemma are satisfied for our choice of matrix A.

Proof. Let v_i be the ith column of A. We show that $v_1,\dots,v_{m-1}$ are linearly independent. Suppose that there exist $t_i \in \mathbb{R}$ with $\sum_{i=1}^{m-1} t_i v_i = 0$, and not all t_i are zero. Choose k such that t_k has maximum absolute value. Dividing the linear dependence relation by t_k, we can assume $t_k = 1$ and all other t_i have absolute value at most 1. Now consider the kth entry of the linear dependence relation.

$$0 = \sum_{i=1}^{m-1} t_i a_{ki} = t_k a_{kk} + \sum_{i \neq k, 1 \leq i \leq m-1} t_i a_{ki}$$

Since $t_i \leq 1, a_{ki} < 0$, we have

$$0 \geq \sum_{i=1}^{m-1} a_{ki} > \sum_{i=1}^{m} a_{ki} \geq 0$$

as $a_{km} < 0$, giving a contradiction as required. □

This proves Dirichlet's unit theorem.

Definition. Let $R_L = \operatorname{covol}(\ell(\mathcal{O}_L^\star) \subseteq \mathbb{R}^{r+s-1})$. This is an invariant of a number field, called the *regulator* of L.

Concretely, choose $\varepsilon_1, \dots, \varepsilon_{r+s-1}$ in $\mathcal{O}_L^\star$ such that $\mathcal{O}_L^\star \simeq \boldsymbol{\mu}_L \times \{\varepsilon_1^{n_1} \dots \varepsilon_{r+s-1}^{n_{r+s-1}} \mid n_i \in \mathbb{Z}\}$. Take any $(r+s-1) \times (r+s-1)$ minor of the $(r+s-1) \times (r+s)$ matrix $(\ell(\varepsilon_1), \dots, \ell(\varepsilon_{r+s}))$. The determinant of the absolute value of this submatrix is R_L.

Example. Let L be a real quadratic field, and let ε be a fundamental unit. Then $\log|\sigma_1(\varepsilon)| = R_L$.

4.3. Finding fundamental units

We now need to find such fundamental units. One way is to guess a unit and then find all smaller ones.

Example. Let $L = \mathbb{Q}(\sqrt{d})$ and $d > 0$, and embed this into $\mathbb{R}$ by choosing $\sqrt{d} > 0$. Consider $d = 2$. One might guess $\varepsilon = 1 + \sqrt{2}$, as $N(\varepsilon) = 1$ so ε is a unit. We claim that this is fundamental. If not, there exists $u = a + b\sqrt{2}$ with $a, b \in \mathbb{Z}$, $u \in \mathcal{O}_L^\star$, and $1 < u < \varepsilon$ as elements of $\mathbb{R}$, identifying L with $\sigma_1(L) \subseteq \mathbb{R}$. The other embedding $\overline{u} = a - b\sqrt{2}$ has $u\overline{u} = \pm 1$. As $u > 1$, $|\overline{u}| < 1$, so $u + \overline{u}, u - \overline{u} > 0$. Hence $a, b > 0$, so there are no possibilities for $1 < a + b\sqrt{2} < 1 + 1\sqrt{2}$ with $a, b > 0$ integers. Hence ε is a fundamental unit.

Example. Consider $d = 11$. Let $\varepsilon = 10 - 3\sqrt{11}$ as $N(\varepsilon) = 1$. Notice that $\varepsilon \approx 0.5$. $\varepsilon^{-1} > 1$ and $\varepsilon^{-1} < 20$. If this were not fundamental, there exists $u = a + b\sqrt{11}$ with $1 < u < \varepsilon^{-1} = 10 + 3\sqrt{11} < 20$. We could check all cases like in the above example, but we can do better in this case. If $N(u) = -1$, we have $a^2 - 11b^2 = -1$, which has no solutions modulo 11 as -1 is not a square in $\mathbb{F}_{11}$. Hence $N(u) = 1$ so $\overline{u} = u^{-1}$, giving $\varepsilon^{-1} > u > 1$ implies $0 < \varepsilon < u^{-1} = \overline{u} < 1$, so $0 < a - b\sqrt{11} < 1$, so $-1 < -a + b\sqrt{11} < 0$. Combining with the previous inequality, $0 < 2b\sqrt{11} < 10 + 3\sqrt{11} < 7\sqrt{11}$ so $b = 1, 2, 3$. Now we can check that $1 + b^2 \cdot 11$ is not a square in $\mathbb{F}_{11}$ for $b = 1, 2, 3$ so there is no possible a. Hence ε is a fundamental unit.

Remark. There is an algorithm for $\mathbb{Q}(\sqrt{d})$ to compute fundamental units. Recall that any real number t can be written as

$$t = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}} = [a_0, a_1, a_2, a_3, \dots]$$

where $a_0 = \lfloor t \rfloor$. t is a quadratic algebraic number, so $[\mathbb{Q}(t) : \mathbb{Q}] = 2$, if and only if the expansion of t as a continued fraction is periodic $t = [a_0, \overline{a_1, \dots, a_m}]$.

The following proposition is non-examinable (and should not be used in exams).

Proposition. Let $\sqrt{d} = [a_0, \overline{a_1, \dots, a_m}]$ and let $\frac{p}{q} = [a_0, \dots, a_{m-1}]$. Then $p + q\sqrt{d}$ is a unit in $L = \mathbb{Q}(\sqrt{d})$, and if $d \equiv 2, 3 \bmod 4$, it is fundamental.

The proof is omitted.

Example. $\sqrt{7} = [2, \overline{1, 1, 1, 4}]$ so $\frac{p}{q} = [2, 1, 1, 1] = \frac{8}{3}$ and $(8 + 3\sqrt{7})(8 - 3\sqrt{7}) = 1$.

This algorithm is polynomial-time in the regulator, but not polynomial-time in the discriminant.

If $q(x, y) = ax^2 + bxy + cy^2$ is a quadratic form for $a, b, c \in \mathbb{Z}$ and $D = b^2 - 4ac$, define $L = \mathbb{Q}(\sqrt{D})$, and define the ideal associated to q to be $\left(a, \frac{-b+\sqrt{D}}{2}\right)$. One can show that if $a > 0, D < 0$, the ideal attached to q is equal to the ideal attached to q' in the class group if and only q and q' are equal under the action of $SL_2(\mathbb{Z})$, i.e. if $q'(x, y) = q(x', y')$

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \underbrace{\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}}_{\in SL_2(\mathbb{Z})} \begin{pmatrix} x \\ y \end{pmatrix}$$

In particular, the size of the class group is exactly the number of orbits of positive definite quadratic forms with discriminant D under the action of $SL_2(\mathbb{Z})$.

5. Dirichlet series and L-functions

5.1. Dirichlet series

Theorem (Euclid)**.** There exist infinitely many primes.

The following proof is due to Euler in 1748.

Proof. Consider

$$\prod_{p \text{ prime}} \left(1 - \frac{1}{p}\right)^{-1} = \prod_{p \text{ prime}} \left(1 + \frac{1}{p} + \frac{1}{p^2} + \dots\right) = \sum_{n=1}^{\infty} \frac{1}{n}$$

as every $n > 0$ factors uniquely as a product of primes so occurs exactly once when we expand the product. If there are finitely many primes, the product is finite. As $\sum_{i=1}^{\infty} p^{-i}$ converges to $\left(1 - \frac{1}{p}\right)^{-1}$, $\sum_{i=1}^{\infty} \frac{1}{n}$ must converge. □

We aim to prove that for all $a, q \in \mathbb{Z}$ coprime, there are infinitely many primes of the form $a + kq$, $k \in \mathbb{N}$. Note that there is no nice series expansion for $\prod_{p \equiv a \bmod q, p \text{ prime}} \left(1 - \frac{1}{p}\right)^{-1}$, so Euler's proof does not generalise.

Definition. The *Riemann zeta function* is $\zeta(s) = \sum_{n \geq 1} n^{-s}$ for $s \in \mathbb{C}$.

Proposition. (i) $\zeta(s)$ converges for $\operatorname{Re}(s) > 1$.

(ii) $\zeta(s) = \prod_{p \text{ prime}} \left(1 - \frac{1}{p^s}\right)^{-1}$ in this region; this result is known as the *Euler product*. This product converges absolutely.

(iii) $\zeta(s) - \frac{1}{s-1}$ extends to a holomorphic function for $\operatorname{Re}(s) > 0$, so the zeta function has a simple pole with residue 1 at $s = 1$.

If the series $\sum \log(1 - a_n)$ converges, $\prod(1 - a_n)$ converges. $\prod(1 - a_n)$ absolutely converges if $\sum |\log(1 - a_n)|$ converges.

If a_n is a sequence of complex numbers, call the function $\sum_{n \geq 1} a_n n^{-s}$ a *Dirichlet series*. Instead of part (i), we will prove the following more general lemma.

Lemma. If there exists $r \in \mathbb{R}$ with $a_1 + \dots + a_N = O(N^r)$, then $\sum_{n \geq 1} a_n n^{-s}$ converges for $\operatorname{Re}(s) > r$, and it is holomorphic in this region.

Proof of lemma.

$$\sum_{n=1}^{N} a_n n^{-s} = a_1(1^{-s} - 2^{-s}) + (a_1 + a_2)(2^{-s} - 3^{-s}) + \dots + (a_1 + a_{N-1})((N-1)^{-s} - N^{-s}) + R_n$$

where $R_n = \frac{T(N)}{N^s}$ with $T(N) = a_1 + \dots + a_N = O(N^r)$. By assumption, if $\mathrm{Re}(s) > r$,

$$\left|\frac{T(N)}{N^s}\right| = \left|\frac{T(N)}{N^r}\right| \cdot \frac{1}{|N^{s-r}|} = \left|\frac{T(N)}{N^r}\right| \cdot \frac{1}{N^{\mathrm{Re}(s)-r}} \to 0$$

as $x^s = e^{s \log x}$ so $|x^s| = |x^{\mathrm{Re}\, s}|$. So if $\mathrm{Re}(s) > r$, $\sum a_n n^{-s} = \sum T(N)(N^{-s} - (N+1)^{-s})$. But $|T(N)| \leq BN^r$ for some constant B by assumption, so it suffices to show $\sum N^r(N^{-s} - (N+1)^{-s})$ converges. Note that

$$N^{-s} - (N+1)^{-s} = \int_N^{N+1} s\frac{\mathrm{d}x}{x^{s+1}}$$

and $N^r \leq x^r$ if $x \in [N, N+1]$. Hence

$$N^r(N^{-s} - (N+1)^{-s}) \leq \int_N^{N+1} x^r s\frac{\mathrm{d}x}{x^{s+1}} \leq s\int_N^{N+1} \frac{\mathrm{d}x}{x^{s+1-r}}$$

It is enough to show that $s\int_1^N \frac{\mathrm{d}x}{x^{s+1-r}}$ converges, which it does to $\frac{s}{s-r}$. □

Proof of proposition. Part (ii). Let $p_1, \dots, p_r$ be the first r primes. Then, $\prod_{i=1}^r (1 - p_r^{-s})^{-1} = \sum_{n \in X} n^{-s}$ where X is the set of positive integers whose prime divisors are only in $p_1, \dots, p_r$. So

$$\left|\zeta(s) - \prod_{i=1}^r (1 - p_r^{-s})^{-1}\right| = \left|\sum_{n \notin X} n^{-s}\right| \leq \sum_{n \notin X} |n^{-s}| = \sum_{n \notin X} n^{-\mathrm{Re}(s)} \leq \sum_{n > r} n^{-\mathrm{Re}(s)}$$

as $1, \dots, r \in X$. Hence the infinite product converges to $\zeta(s)$. The proof of absolute convergence is omitted.

Part (iii). Left as an exercise, noting that

$$\frac{1}{s-1} = \sum_{i=1}^{\infty} \int_n^{n+1} \frac{\mathrm{d}t}{t^s}$$

□

5.2. Zeta functions in number fields

The remaining new content in this course is nonexaminable.

Definition. Let L be a number field. The *zeta function of L* is

$$\zeta_L(s) = \sum_{\mathfrak{a} \trianglelefteq \mathcal{O}_L} N(\mathfrak{a})^{-s} = \sum_{n \geq 1} \#\{\mathfrak{a} \trianglelefteq \mathcal{O}_L \mid N(\mathfrak{a}) = n\} n^{-s}$$

Proposition. (i) $\zeta_L(s)$ converges to a holomorphic function for $\mathrm{Re}(s) > 1$.

(ii) $\zeta_L(s) = \prod_{\mathfrak{p} \text{ prime ideal}} (1 - N(\mathfrak{p})^{-s})^{-1}$ in this region.

(iii) $\zeta_L(s)$ is a meromorphic function for $\mathrm{Re}(s) > 1 - \frac{1}{[L:\mathbb{Q}]}$, with a simple pole at $s = 1$ with residue
$$\frac{|\mathrm{Cl}_L| 2^{r+s} \pi^s R_L}{|D_L|^{\frac{1}{2}} |\boldsymbol{\mu}_L|}$$
This is called the *analytic class number formula.*

Proof. Part (ii) is clear. Parts (i) and (iii) follow from the following estimate. Writing $\zeta_L(s) = \sum \frac{a_n}{n^s}$ where a_n is the number of ideals of norm n, one can show

$$a_1 + \cdots + a_N = \frac{|\mathrm{Cl}_L| 2^{r+s} \pi^s R_L}{|D_L|^{\frac{1}{2}} |\boldsymbol{\mu}_L|} \cdot N + O\left(N^{1 - \frac{1}{[L:\mathbb{Q}]}}\right)$$

□

If $L \neq \mathbb{Q}$, it turns out that $\zeta_L(s)$ factors into $\zeta_{\mathbb{Q}}(s) = \zeta(s)$ and some other factors. Suppose $L = \mathbb{Q}(\sqrt{d})$ and $d \neq 0, 1$ is square-free.

$$\zeta_L - \prod_{\mathfrak{p} \text{ prime ideal}} (1 - N(\mathfrak{p})^{-s})^{-1} = \prod_{p \text{ prime}} \prod_{\mathfrak{p} | (p)} (1 - N(\mathfrak{p})^{-s})^{-1}$$

If $p \mid D_L$, then $(p) = \mathfrak{p}^2$ ramifies. In this case, $N(\mathfrak{p}) = p$ and we have a term $(1 - p^{-s})$ in the product. If (p) remains prime in L, then $N(\mathfrak{p}) = p^2$ giving the term $(1 - p^{-2s}) = (1 - p^{-s})(1 - p^s)$. If $(p) = \mathfrak{p}_1 \mathfrak{p}_2$ splits, then $N(\mathfrak{p}_i) = p$ and we have a term $(1 - p^{-s})^2$. Let

$$\chi_{D_L}(p) = \chi(p) = \begin{cases} 0 & p \text{ ramifies} \\ -1 & p \text{ inert} \\ 1 & p \text{ splits} \end{cases} = \underbrace{\left(\frac{D_L}{p}\right)}_{\text{if } p \text{ odd}}$$

Then, defining $L(\chi, s) = \prod_{p \text{ prime}} 1 - \chi(p) p^{-s}{}^{-1}$, we have $\zeta_L(s) = \zeta_{\mathbb{Q}}(s) L(\chi, s)$. The function L is called a *Dirichlet L-function.* When expanding the infinite product defining $L(\chi_D, s)$ the coefficient of n^{-s}, if $n = p_1^{e_1} \dots p_r^{e_r}$ is $\chi_D(p_1)^{e_1} \dots \chi_D(p_r)^{e_r}$. We can extend the definition of χ to make it multiplicative: $\chi_D(p_1^{e_1} \dots p_r^{e_r}) = \chi_D(p_1)^{e_1} \dots \chi_D(p_r)^{e_r}$.

Example. Let $L = \mathbb{Q}(\sqrt{-1})$, so $D_L = 4$. We have $\left(\frac{-4}{p}\right) = \left(\frac{-1}{p}\right) = (-1)^{\frac{p-1}{2}}$ for $p \neq 2$. 2 ramifies, so $\chi_D(2) = 0$. We claim that

$$\chi_{-4}(m) = \begin{cases} (-1)^{\frac{m-1}{2}} & m \text{ odd} \\ 0 & m \text{ even} \end{cases}$$

Indeed, if n is even, this is clear; otherwise, this claim is that $(-1)^{\frac{mn-1}{2}} = (-1)^{\frac{m-1}{2}} (-1)^{\frac{n-1}{2}}$, which is easy to verify. Hence,

$$L(\chi_{-4}, s) = 1 - \frac{1}{3^s} + \frac{1}{5^s} - \frac{1}{7^s} + \dots$$

In this example, the coefficients are periodic mod 4; this is true for general $L(\chi_D, s)$. Since $\zeta_L(s) = \zeta_{\mathbb{Q}}(s)L(\chi_{-4}, s)$, the fact that $\zeta_{\mathbb{Q}}(s)$ has a simple pole at $s = 1$ with residue 1, together with the analytic class number formula, gives $L(\chi_{-4}, 1) = \frac{\pi}{4}$.

Definition. $\chi : \mathbb{Z} \to \mathbb{C}$ is a *Dirichlet character* of modulus D if there exists a group homomorphism $\omega : \left(\mathbb{Z}/D\mathbb{Z}\right)^\star \to \mathbb{C}$ such that

$$\chi(n) = \begin{cases} \omega(n \bmod D) & n \text{ invertible mod } D \\ 0 & \text{otherwise} \end{cases}$$

For such a χ, we have $\chi(n)\chi(m) = \chi(nm)$, and we can define

$$L(\chi, s) = \prod_{p \text{ prime}} (1 - \chi(p)p^{-s})^{-1} = \sum_{n \geq 1} \chi(n)n^{-s}$$

The previous example shows that χ_{-4} is a Dirichlet character of modulus 4.

Theorem. For any $d \neq 0, 1$ square-free, defining $L = \mathbb{Q}(\sqrt{d}), D = D_L$, we have that χ_D is a Dirichlet character of modulus D.

Proof. We must show $\chi_D(n + D) = \chi_D(n)$ for $n \in \mathbb{N}$. Suppose first that $d \equiv 3 \bmod 4$. Here, $D = 4d$, so $\chi_D(2) = 0$ as 2 ramifies, so $\chi_D(n) = 0$ if n is even as required. For $p > 2$, $\chi_D(p) = \left(\frac{D}{p}\right) = \left(\frac{d}{p}\right)$ by definition, but this is equal to $\left(\frac{p}{d}\right)(-1)^{\frac{p-1}{2}}$ by quadratic reciprocity as p, d are odd, and as $d \equiv 3 \bmod 4$, $\frac{d-1}{2} \equiv 1 \bmod 4$. $n \mapsto (-1)^{\frac{n-1}{2}}$ is multiplicative, so $\chi_D(n + D) = \left(\frac{n+D}{d}\right)(-1)^{\frac{n-1}{2}}(-1)^{4d}2 = \chi_D(n)$. The other cases are omitted. □

This theorem can be seen as equivalent to the law of quadratic reciprocity. Note that χ is nontrivial if $\omega \not\equiv 1$

Lemma. If χ is a nontrivial Dirichlet character, $L(\chi, s)$ is holomorphic for $\operatorname{Re} s > 0$.

Proof. Recall that if G is a finite group and χ_1, χ_2 are characters of irreducible complex representations, then

$$\frac{1}{G} \sum_{g \in G} \overline{\chi_1(g)}\chi_2(g) = \begin{cases} 1 & \chi_1 = \chi_2 \\ 0 & \text{otherwise} \end{cases}$$

Applying this to $G = \left(\mathbb{Z}/d\mathbb{Z}\right)^\star$ where χ_1 is the trivial character and $\chi_2 = \omega$, this gives

$$\sum_{ad < i < (a+1)d} \chi(i) = \sum_{i \in \mathbb{Z}/d\mathbb{Z}} \chi(i) = \sum_{i \in \left(\mathbb{Z}/d\mathbb{Z}\right)^\star} \omega(i) = 0$$

In particular, $\sum_{i=1}^n \chi(i) = O(1)$ is bounded. So $\sum_{i=1}^n \frac{\chi(i)}{n^s}$ converges for $\operatorname{Re}(s) > 0$. □

Corollary. If $D < 0$,

$$L(\chi_D, 1) = \frac{2\pi\left|\mathrm{Cl}_{\mathbb{Q}(\sqrt{D})}\right|}{|D|^{\frac{1}{2}}\left|\mu_{\mathbb{Q}(\sqrt{D})}\right|}$$

In particular, $L(\chi_D, 1) \neq 0$.

Proof. $\zeta_{\mathbb{Q}(\sqrt{(D)})}(s) = \zeta_{\mathbb{Q}}(s)L(\chi_D, s)$, so both sides have a simple pole at $s = 1$. The analytic class number formula gives the residue of the left hand side, and $\mathrm{Res}_\zeta(1) = 1$. □

5.3. L-functions in cyclotomic fields

We will show that $L(\chi, 1) \neq 0$ for any Dirichlet character χ, and hence show that there are infinitely many primes in arithmetic progression. To do this, we will factor $\zeta_{\mathbb{Q}(e^{\frac{2\pi i}{q}})}$ for any q. Consider $L = \mathbb{Q}(\omega_q)$ where ω_q is a primitive qth root of unity,

Proposition. (i) $[L : \mathbb{Q}] = \varphi(q)$ where $\varphi(q) = \left|\left(\mathbb{Z}/q\mathbb{Z}\right)^\star\right|$;

(ii) $L/\mathbb{Q}$ is a Galois extension with Galois group $G = \left(\mathbb{Z}/q\mathbb{Z}\right)^\star$, and if $r \in \left(\mathbb{Z}/q\mathbb{Z}\right)^\star$, then r acts on L by mapping ω_q to ω_q^r;

(iii) $\mathcal{O}_L = \mathbb{Z}[\omega_q] = \mathbb{Z}[x]/\Phi_q(x)$ where Φ_q is the qth cyclotomic polynomial;

(iv) if p is prime, $p \mid D_L$ if and only if $p \mid q$;

(v) if p is prime, p ramifies in $\mathcal{O}_L$ if and only if $p \mid q$;

(vi) if p is prime with $p \nmid q$, then (p) factors as a product of $\frac{\varphi(q)}{f}$ distinct prime ideals, each of norm p^f, where f is the order of p in $\left(\mathbb{Z}/q\mathbb{Z}\right)^\star$.

Proof. Parts (i) and (ii) follow from Galois theory. Part (iii) for q prime is on an example sheet, and the general case is omitted. Part (iv) is omitted. Part (iv) implies (v) is a general fact; we will only show part (vi).

As $\mathcal{O}_L = \mathbb{Z}[x]/\Phi_q(x)$, Dedekind's theorem applies. We study $\mathcal{O}_L/(p) = \mathbb{F}_p[x]/\Phi_q(x)$ by factoring $\Phi_q(x)$ modulo p. Recall that

$$\Phi_q(x) = \frac{x^q - 1}{\prod_{d\neq q, d\mid q} \Phi_d(x)}$$

so for instance $\Phi_8(x) = \frac{x^8-1}{x^4-1} = x^4 + 1$.

$$\left(\mathbb{Z}/8\mathbb{Z}\right)^\star = \{1, 3, -3, -1\} \simeq \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$$

In this example, if $p = 17$, $x^4 + 1$ factors into four linear factors, but if $p = 3$, $x^4 + 1$ factors into two factors as the order of 3 is 2 in $\left(\mathbb{Z}/8\mathbb{Z}\right)^\star$.

Write $\Phi_q(x) = \gamma_1^{e_1} \dots \gamma_g^{e_g}$ for γ_i irreducible and distinct, so

$$\mathcal{O}_L/(p) = \mathbb{F}_p[x]/(\gamma_1^{e_1}) \times \dots \mathbb{F}_p[x]/(\gamma_g^{e_g})$$

For any number field L, $\mathrm{Gal}(L/\mathbb{Q})$ preserves $\mathcal{O}_L$. Indeed, if $\alpha \in \mathcal{O}_L$, $f(\alpha) = 0$ for some monic polynomial $f \in \mathbb{Z}[x]$, but then $g \in \mathrm{Gal}(L/\mathbb{Q})$ givesn $0 = gf(\alpha) = f(g(\alpha)) = 0$, so $g(\alpha)$ is also a root of f and hence in $\mathcal{O}_L$.

G permutes the roots of Φ_q, so G acts on $\{\gamma_1, \dots, \gamma_g\}$. This action is transitive on the roots, so is transitive on $\{\gamma_1, \dots, \gamma_g\}$. Hence $\deg \gamma_1 = \dots = \deg \gamma_g$, so $e_1 = \dots = e_g = e$. Further, ge is the order of $G/\mathrm{Stab}_G(\gamma_1)$.

If $p \nmid D_L$, or equivalently $p \nmid q$, then $e = 1$ as p is unramified. Hence $\mathbb{F}_p[x]/(\gamma_1) = \mathbb{F}_{p^{f'}}$ for some f', and $\frac{\varphi(q)}{f'}$ factors. We must show that $f' = f$.

$p \in \left(\mathbb{Z}/q\mathbb{Z}\right)^\star = \mathrm{Gal}(L/\mathbb{Q})$ acts as $\alpha \mapsto \alpha^p$ on $\mathbb{F}_{p^{f'}}$, so it acts as the Frobenius automorphism, which is the generator of the Galois group of $\mathbb{F}_{p^{f'}}/\mathbb{F}_p$ by (ii). Conversely, the image of x in $\mathbb{F}_p[x]/(\gamma_1)$, is the image of ω_q which is a primitive qth root of unity. So $q \mid \left|\mathbb{F}^\star_{p^{f'}}\right|$, so $q \mid p^{f'} - 1$. In particular, $p^{f'} \equiv 1 \bmod q$, so $f = \mathrm{ord}(p) \mid f'$. Hence $f = f'$ as required. □

Recall that $\zeta_{\mathbb{Q}(\omega_q)}(s) = \prod_{\mathfrak{p}\text{ prime}} (1 - N(\mathfrak{p})^{-s})^{-1}$. Consider prime ideals $\mathfrak{p}$ dividing (p) for a fixed integer prime p. If $p \nmid q$, part (vi) shows that these contribute $(1 - p^{-fs})^{-\frac{\varphi(q)}{f}}$ to the zeta function, where f is the order of p in $\left(\mathbb{Z}/q\mathbb{Z}\right)^\star$. But this factors as $(1 - t^f) = \prod_{\gamma \in \mu_f} (1 - \gamma t)$ where $\mu_f = \{\gamma \in \mathbb{C} \mid \gamma^f = 1\}$.

Set $t = p^{-s}$, and let $\omega_1, \dots, \omega_{\varphi(q)} : \left(\mathbb{Z}/q\mathbb{Z}\right)^\star \to \mathbb{C}$ be the distinct irreducible complex representations of $\left(\mathbb{Z}/q\mathbb{Z}\right)^\star$, such that $\omega_1 = \mathbb{1}$ so $\omega_1(\alpha) = 1$ for all $\alpha \in \left(\mathbb{Z}/q\mathbb{Z}\right)^\star$. We claim that $\omega_1(p), \dots, \omega_{\varphi(q)}(p)$ are the distinct fth roots of unity, each repeated $\frac{\varphi(q)}{f}$ times. Certainly p generates a cyclic subgroup (p) of $\left(\mathbb{Z}/q\mathbb{Z}\right)^\star$ of order f by definition of f. The claim is that the restriction of of $\omega_1, \dots, \omega_{\varphi(q)}$ to (p) are the f distinct irreducible representations of (p), each repeated $\frac{\varphi(q)}{f}$ times, which can be easily proven using representation theory. We have therefore shown that

$$(1 - p^{-fs})^{-\frac{\varphi(q)}{f}} = \prod_{i=1}^{\varphi(q)} (1 - \omega_i(p) p^{-s})^{-1}$$

Let

$$\chi_i(n) = \begin{cases} \omega_i(n \bmod q) & \text{if } \gcd(n,q) = 1 \\ 0 & \text{otherwise} \end{cases}$$

Then we have shown that

$$\zeta_{\mathbb{Q}(\omega_q)}(s) = \prod_{i=1}^{\varphi(q)} L(\chi_i, s) \text{ multiplied by a correction term}$$

which is a finite product of the form $\prod_{p \mid q}(1 - p^{-f_p s})^{-1}$. Note that $\zeta_{\mathbb{Q}}(s) = L(\chi_1, s)\prod_{p \mid q}(1 - p^{-s})^{-1}$, so we can rewrite this as

$$\zeta_{\mathbb{Q}(\omega_p)}(s) = \zeta_{\mathbb{Q}}(s)\prod_{i=2}^{\varphi(q)} L(\chi_i, s) \text{ multiplied by a correction term}$$

Theorem. If χ_i is a nontrivial Dirichlet character, then $L(\chi_i, 1) \neq 0$.

In fact, if χ is any nontrivial Dirichlet character modulo q, $\chi = \chi_i$ for some i.

Proof. We have shown that if χ is a nontrivial Dirichlet character, $L(\chi, s)$ is holomorphic at $s = 1$. In the above expansion, the left hand side and right hand side are meromorphic functions at $s = 1$ with a simple pole. The residue of the right hand side and left hand side therefore agree, and its value is

$$\operatorname{Res}_{s=1} \zeta_{\mathbb{Q}}(s)\prod_{i=2}^{\varphi(q)} L(\chi_i, 1) \text{ multiplied by a correction term}$$

The analytic class number formula implies that this is nonzero, so $L(\chi_i, 1) \neq 0$. □

Note that Dirichlet characters of quadratic fields have values in ±1.

5.4. Primes in arithmetic progression

Theorem (Dirichlet). Let $a, q \in \mathbb{N}$ with $\gcd(a, q) = 1$. There are infinitely many primes in $a, a + q, a + 2q, \dots$.

Proof. Consider $\left(\mathbb{Z}/q\mathbb{Z}\right)^\star$, an abelian group of order $\varphi(q)$. Let $\omega_1, \dots, \omega_{\varphi(q)} : \left(\mathbb{Z}/q\mathbb{Z}\right)^\star \to \mathbb{C}^\star$ where $\omega_1 = \mathbb{1}$, and $\chi_1, \dots, \chi_{\varphi(q)} : \mathbb{Z}/q\mathbb{Z} \to \mathbb{C}$ be the corresponding Dirichlet characters. Recall the orthogonality of the columns of the character table of a finite group:

$$\frac{1}{\varphi(q)}\sum_i \overline{\omega_i(a)}\omega_i(p) = \begin{cases} 1 & a \equiv p \mod q \\ 0 & \text{otherwise} \end{cases}$$

if $\gcd(p, q) = 1$, so p defines an element of $\left(\mathbb{Z}/q\mathbb{Z}\right)^\star$. Hence,

$$\frac{1}{\varphi(q)} \sum_i \overline{\chi_i(a)}\chi_i(p) = \begin{cases} 1 & a \equiv p \mod q \\ 0 & \text{otherwise} \end{cases}$$

even if $\gcd(p, q) \neq 1$, since in this case $\chi_i(p) = 0$ by definition. Hence,

$$\sum_{p \equiv a \text{ mod } q, p \text{ prime}} p^{-s} = \frac{1}{\varphi(q)} \sum_{i,p} \overline{\chi_i(a)}\chi_i(p)p^{-s}$$

We want to show that this has a pole at $s = 1$. If χ is a Dirichlet character, by the series expansion of logarithm which is valid by absolute convergence, we can write

$$\begin{aligned} \log L(\chi, s) &= -\sum_p \log(1 - \chi(p)p^{-s}) \\ &= \sum_{n,p} \frac{\chi(p)^n}{p^{ns}n} \\ &= \sum_{n,p} \frac{\chi(p^n)}{p^{ns}n} \\ &= \sum_p \frac{\chi(p)}{p^s} + \sum_{n \geq 2, p} \frac{\chi(p^n)}{p^{ns}n} \end{aligned}$$

We claim that $\sum_{n \geq 2, p \text{ prime}} \frac{\chi(p^n)}{p^{ns}n}$ converges at $s = 1$. This holds as its absolute value is at most

$$\sum_{n \geq 2, p \text{ prime}} p^{-ns} = \sum_{p \text{ prime}} \frac{1}{p^s(p^s - 1)} \leq \sum_n \frac{1}{n^s(n^s - 1)} \leq 2\frac{1}{n^{2s}}$$

which is finite at $s = 1$. Hence, the series above has a pole at $s = 1$ if and only if

$$\frac{1}{\varphi(q)} \sum_i \overline{\chi_i(a)} \log L(\chi_i, s)$$

has a pole at $s = 1$.

If χ_1 is the trivial character, $L(\chi_1, s) = \zeta_{\mathbb{Q}}(s) \prod_{p|s}(1 - p^{-s})$, so as $\zeta_{\mathbb{Q}}(s)$ has only a simple pole at $s = 1$, $\log \zeta_{\mathbb{Q}}(s) = \log \frac{1}{s-1} +$ bounded function near $s = 1$, so $\log L(\chi_1, s) \sim \log \frac{1}{s-1}$ has a pole at $s = 1$. For $i \neq 1$, $L(\chi_i, s)$ is nonzero at $s = 1$ by the above theorem, so $\log L(\chi_i, s)$ is bounded at $s = 1$. Hence, $\frac{1}{\varphi(q)} \sum_{i,p} \overline{\chi_i(a)}\chi_i(p)p^{-s} \sim \frac{1}{\varphi(q)} \log \frac{1}{s-1}$, and in particular has a pole at $s = 1$.

Hence, there are infinitely many primes in arithmetic progression. □

This proof shows that approximately $\frac{1}{\varphi(q)}$ of all primes lie in this arithmetic progression.

One can in fact show that for any number field L, $\zeta_L(s)$ always factors and the factors have meaning. Suppose $L/\mathbb{Q}$ is Galois, and $G = \text{Gal}(L/\mathbb{Q})$. Then,

(i) We can factor $\zeta_L(s) = \prod_{\rho \text{ irreducible representation of } G} L(\rho, s)^{\dim \rho}$, where the $L(\rho, s)$ are *Artin L-functions*. Moreover, $L/\mathbb{Q}$ is the regular representation of G.

(ii) $L(\mathbb{1}, s) = \zeta_{\mathbb{Q}}(s)$.

(iii) $L(\rho, s)$ is a meromorphic function of s. It is conjectured, but still not known, that $L(\rho, s)$ is holomorphic if $\rho \neq \mathbb{1}$.

(iv) If ρ is one-dimensional, then $L(\rho, s) = L(\chi, s)$ multiplied by a correction factor, where $L(\chi, s)$ is a Dirichlet L-function. Finding χ given ρ is a generalisation of quadratic reciprocity, called class field theory.

(v) The properties of mutidimensional ρ are studied in the Langlands programme.

IX. Algebraic Geometry

Lectured in Lent 2023 by Dr. D. Ranganathan

In this course, we study the duality between systems of polynomial equations and the geometry or topology of their solution sets. Sets of points that arise as solution sets of polynomials are called algebraic varieties. We therefore study the correspondence between sets of polynomials and the varieties they define.

One can show that the set of varieties satisfy the axioms of the closed sets of a topological space. We can thus define a topology where the closed sets are precisely the algebraic varieties; this is called the Zariski topology. This very explicit description of the closed sets allows us to study this topology in depth. There are some drawbacks; the Zariski topology is not even Hausdorff.

Some geometric properties of algebraic varieties can be studied algebraically. For example, the dimension of a variety is the amount of algebraically independent transcendental elements of a field associated to the variety. Perhaps the simplest varieties are the curves, those varieties with dimension 1. They have comparatively simple field structure, and are studied in depth.

Contents

1. Affine varieties

1.1. Introduction

Algebraic geometry studies the duality between systems of polynomial equations and the geometry or topology of their solution sets. If we have a system of polynomials

$$f_1, \dots, f_r \in \Bbbk[X_1, \dots, X_n] = \Bbbk[\mathbf{X}]$$

we can form its solution set

$$V = \{P \in \Bbbk^n \mid f_1(P) = \dots = f_r(P) = 0\} \subseteq \Bbbk^n$$

On the algebraic side, we have the ideal

$$I = (f_1, \dots, f_r) \lhd \Bbbk[\mathbf{X}]$$

The duality we are interested in is between $R = {}^{\Bbbk[\mathbf{X}]}\!/_{I}$ and the geometry of V.

We may impose some assumptions on the field $\Bbbk$.

- We might assume that $\Bbbk$ is algebraically closed, which is a natural assumption since we wish to consider roots to polynomials with coefficients in $\Bbbk$.
- We could also take the stronger assumption that $\Bbbk$ is algebraically closed and has characteristic 0. Occasionally, we may want to differentiate a polynomial, and so it becomes inconvenient to do algebra without this assumption.
- Throughout the course, we will in fact assume $\Bbbk = \mathbb{C}$, as we are not particularly interested in the subtleties of such fields other than $\mathbb{C}$, and it is useful for intuition.

Questions we may ask about this duality are:

- To what extent do R and V determine each other?
- What is the right notion of dimension of V, in terms of algebra?
- Can we detect whether $V \subseteq \mathbb{C}^n$ is a manifold based on the information contained within R?
- Is V compact? If not, is there a natural way to compactify the space into some space $\overline{V}$ that is in some sense algebraic?

1.2. Affine space

Definition. The *affine space of dimension n*, implicitly over $\mathbb{C}$, is the set $\mathbb{A}^n = \mathbb{C}^n$. The elements of $\mathbb{A}^n$ are called *points*, denoted $P = (\mathbf{a}) = (a_1, \dots, a_n)$.

Definition. An *affine subspace* of $\mathbb{A}^n$ is any subset of the form $v + U \subseteq \mathbb{C}^n$ where $U \subseteq \mathbb{C}^n$ is any linear subspace, and $v \in \mathbb{C}^n$.

$\mathbb{A}^n$ is the natural set on which $\mathbb{C}[X_1, \dots, X_n]$ is a ring of functions. Given $f \in \mathbb{C}[\mathbf{X}]$, we obtain a function $f : \mathbb{A}^n \to \mathbb{C}$. The subset $\mathbb{C} \subseteq \mathbb{C}[\mathbf{X}]$ is the set of constant functions.

Proposition. The polynomial ring $\mathbb{C}[\mathbf{X}]$ satisfies the following properties.

(i) $\mathbb{C}[\mathbf{X}]$ is a unique factorisation domain.

(ii) Every ideal in $\mathbb{C}[\mathbf{X}]$ is finitely generated (equivalently, $\mathbb{C}[\mathbf{X}]$ is Noetherian), due to the Hilbert basis theorem.

1.3. Affine varieties

Definition. Let $S \subseteq \mathbb{C}[\mathbf{X}]$ be any subset of $\mathbb{C}[\mathbf{X}]$. The *vanishing locus* of S is defined to be $\mathbb{V}(S) = \{P \in \mathbb{A}^n \mid \forall f \in S,\ f(P) = 0\}$.

Definition. An *affine (algebraic) variety in* $\mathbb{A}^n$ is a set of the form $\mathbb{V}(S)$ for some S.

Note that there is some inconsistency between definitions in different textbooks; some authors also impose an irreducibility condition.

Example. (i) Let $n = 1$. The polynomial $f \in \mathbb{C}[X]$ gives the vanishing locus $\mathbb{V}(f) \subseteq \mathbb{A}^1$, the set of zeroes of f. Conversely, if $V \subseteq \mathbb{A}^1$ is finite, then $V = \mathbb{V}(f)$ where $f = \prod_{a \in V}(x - a)$.

(ii) A *hypersurface* in $\mathbb{A}^n$ is a variety of the form $\mathbb{V}(f)$ where $f \in \mathbb{C}[X]$.

(iii) It is often convenient to represent varieties not by equations but parametrically. The *affine twisted cubic* is $C = \left\{(t, t^2, t^3) \mid t \in \mathbb{C}\right\} \subset \mathbb{A}^3$. This is a variety, as it is the vanishing locus of the two polynomials $X_1^2 - X_2$ and $X_1^3 - X_3$.

Theorem. Let $S \subseteq \mathbb{C}[\mathbf{X}]$. Then,

(i) Let $I \subseteq \mathbb{C}[\mathbf{X}]$ be the ideal generated by S. Then, $\mathbb{V}(S) = \mathbb{V}(I)$.

(ii) There exists a finite subset $\{f_j\}$ of S such that $\mathbb{V}(S) = \mathbb{V}(\{f_j\})$.

Proof. Part (i). Suppose $P \in \mathbb{A}^n$. Then, $f(P) = 0$ for all $f \in S$ if and only if $f(P) = 0$ for all $f \in I$, by the basic properties of ideals.

Part (ii). By (i), $\mathbb{V}(S) = \mathbb{V}(I)$. I is finitely generated, so there exist functions $h_1, \dots, h_r \in I$ that generate I. Reversing (i), $\mathbb{V}(I) = \mathbb{V}(\{h_i\})$. But since I is generated by S, each h_i can be written as a linear combination of finitely many elements of S. So $h_i = \sum_j g_{ij} f_j$ where $f_j \in S$. Then $\mathbb{V}(S) = \mathbb{V}(\{f_j\})$. □

Proposition. Let $S, T \subseteq \mathbb{C}[\mathbf{X}]$. Then,

(i) $S \subseteq T$ implies $\mathbb{V}(T) \subseteq \mathbb{V}(S)$.

(ii) $\mathbb{V}(0) = \mathbb{A}^n$, and $\mathbb{V}(\mathbb{C}[\mathbf{X}]) = \mathbb{V}(\lambda) = \varnothing$ where $\lambda \in \mathbb{C} \setminus \{0\}$.

(iii) $\bigcap_j \mathbb{V}(I_j) = \mathbb{V}\left(\sum_j I_j\right)$ for any family of ideals I_j.

(iv) $\mathbb{V}(I) \cup \mathbb{V}(J) = \mathbb{V}(I \cap J)$.

Proof. Part (i) and (ii) are trivial.

Part (iii). We have $\bigcap_j \mathbb{V}(I_j) = \mathbb{V}\left(\bigcup_j I_j\right)$. To conclude, note that the ideal generated by $\bigcup_j I_j$ is $\sum_j I_j$.

Part (iv). We have already seen that $\mathbb{V}(I) \cup \mathbb{V}(J) \subseteq \mathbb{V}(I \cap J)$. For the reverse containment, suppose $P \in \mathbb{V}(I \cap J)$, and suppose $P \notin \mathbb{V}(I)$. Then, there exists some $g \in I$ such that $g(P) = 0$. Moreover, for all elements $f \in J$, $fg \in I \cap J$, so $(fg)(P) = 0$. Hence $f(P) = 0$ for all $f \in J$, so $P \in \mathbb{V}(J)$. □

1.4. Irreducible varieties

Definition. A variety V is called *irreducible* if whenever $V = V_1 \cup V_2$, where V_1, V_2 are varieties, we have $V = V_1$ or $V = V_2$. A variety that is not irreducible is called reducible.

Example. The variety $V = \mathbb{V}(XY)$ is reducible, as it is the union of $\mathbb{V}(X)$ and $\mathbb{V}(Y)$.

Proposition. Every affine variety V is a finite union of irreducible varieties.

This proof uses a 'bisection' argument.

Proof. If V is irreducible, there is nothing to prove. Otherwise, $V = V_1 \cup V_1'$, where $V_1, V_1' \neq V$. If V_1, V_1' are finite unions of irreducible varieties, the proof is already complete. Suppose V_1 is not a finite union of irreducibles. Then, it follows that $V_1 = V_2 \cup V_2'$ nontrivially. Inductively, we obtain

$$V = V_0 \supsetneq V_1 \supsetneq V_2 \supsetneq V_3 \supsetneq \dots$$

This infinite descending chain never stabilises. Define

$$W = \bigcap_{j=0}^{\infty} V_j = \mathbb{V}\left(\sum_{j=0}^{\infty} I_j\right)$$

But $\sum_{j=0}^{\infty} I_j$ is finitely generated. So $\sum_{j=0}^{\infty} I_j = \sum_{j \leq N} I_j$ for some $N \in \mathbb{N}$. Hence, $W = \bigcap_{j \leq N} V_j$ contradicting that the descending chain never stabilises. □

Definition. Let V be an affine variety. A *minimal decomposition* of V is a representation of V as a finite union of distinct irreducibles V_i such that no V_i is contained within V_j.

Proposition. Minimal decompositions of affine varieties are unique up to ordering.

Proof sketch. This proof is left as an exercise. One can compare two decompositions by intersecting the irreducible components of one decomposition with the other. □

Given uniqueness of minimal decompositions, we can refer to the irreducibles appearing in such a decomposition as the *irreducible components* of a variety.

1.5. Zariski and Euclidean topologies

Definition. The *Zariski topology* on $\mathbb{A}^n$ is the topology where the closed sets are precisely the affine varieties. If $V \subseteq \mathbb{A}^n$ is a (sub)variety, the Zariski topology on V is the subspace topology for the Zariski topology on $\mathbb{A}^n$.

Remark. This is in fact a topology, as all of the relevant axioms have been proven.

Definition. The *Euclidean topology* or *analytic topology* on $\mathbb{A}^n$ is the topology induced by the metric space structure on $\mathbb{C}^n$. If $V \subseteq \mathbb{A}^n$, the Euclidean topology on V is the subspace topology of the Euclidean topology on $\mathbb{A}^n$.

Proposition. The Zariski topology on $\mathbb{A}^1$ coincides with the cofinite topology; the closed sets are exactly the finite sets. This topology is not Hausdorff but it is compact. The Euclidean topology on $\mathbb{A}^1$ is Hausdorff but not compact.

Remark. $\mathbb{A}^2$ with the Zariski topology is not homeomorphic to $\mathbb{A}^1 \times \mathbb{A}^1$ with the product of the Zariski topologies.

1.6. Ideals from zero sets

Theorem (weak form of Hilbert's Nullstellensatz)**.** Every maximal ideal in $\mathbb{C}[\mathbf{X}]$ has the form $(X_1 - a_1, \dots, X_n - a_n)$ for $a_i \in \mathbb{C}$. Moreover, if I is any non-unit ideal, $\mathbb{V}(I) \neq \varnothing \subseteq \mathbb{A}^n$.

We prove this over the complex numbers; the given proof only works for this case, but the statement holds for all algebraically closed fields.

Proof. Every ideal of this form has quotient $\mathbb{C}$, so they are all maximal. Let $\mathfrak{m} \lhd \mathbb{C}[\mathbf{X}]$ be a maximal ideal, and let $K = {}^{\mathbb{C}[\mathbf{X}]}\!/\!_{\mathfrak{m}}$. K is a field as $\mathfrak{m}$ is maximal, and it is a field extension of $\mathbb{C}$. Define a_i to be the coset $X_i + \mathfrak{m}$. If $a_i \in \mathbb{C}$ for all i, this gives the result as required because the ideal is generated by $(X_1 - a_1, \dots, X_n - a_n)$.

Otherwise, $K \supsetneq \mathbb{C}$. But $\mathbb{C}$ is algebraically closed, so there exists $t \in K \setminus \mathbb{C}$ which is transcendental over $\mathbb{C}$. Let U_m be the $\mathbb{C}$-span inside K of products of the form $a_1^{r_1} \dots a_n^{r_n}$ where the r_i are nonnegative, and $\sum_{i=1}^n r_i \leq m$. Observe that U_m is finite-dimensional, and $K = \bigcup_{m \geq 0} U_m$ is countable-dimensional. One can show that the elements $\left\{ \frac{1}{t-c} \mid c \in \mathbb{C} \right\}$ are linearly independent over $\mathbb{C}$. There are uncountably many such elements, giving a contradiction.

For the last part, let I be a nonzero ideal. There exists a maximal ideal $\mathfrak{m} \supseteq I$, so $\mathbb{V}(I) \supseteq \mathbb{V}(\mathfrak{m})$, but $\mathbb{V}(\mathfrak{m})$ is nonempty as it contains the point $(a_1, \dots, a_m)$. □

Definition. Let $V \subseteq \mathbb{A}^n$ be an affine variety. The *ideal of functions vanishing on* V is $I(V) = \{f \in \mathbb{C}[\mathbf{X}] \mid \forall P \in V,\ f(P) = 0\}$.

Proposition. Let $V \subseteq \mathbb{A}^n$ be an affine variety. Then,

(i) If $V = \mathbb{V}(S)$ where $S \subseteq \mathbb{C}[\mathbf{X}]$, then $S \subseteq I(V)$. In particular, $I(V)$ is the largest ideal vanishing on V.

(ii) $V = \mathbb{V}(I(V))$.

(iii) Varieties $V, W \subseteq \mathbb{A}^n$ are equal if and only if $I(V) = I(W)$.

Proof. Follows from the definitions. □

Therefore, we have an injective map I from the space of affine varieties in $\mathbb{A}^n$ to the space of ideals in $\mathbb{C}[\mathbf{X}]$, and $\mathbb{V}$ gives a left inverse.

Proposition. If V, W are affine varieties, $V \subseteq W$ if and only if $I(W) \subseteq I(V)$.

Proof. The forward implication follows from set theory. For the reverse, if $V \not\subseteq W$, we can choose $P \in V \setminus W$. Since $P \notin \mathbb{V}(I(W))$, there exists a function $f \in I(W)$ such that $f(P) \neq 0$, so $f \notin I(V)$. □

Proposition. Let V be a variety. Then V is irreducible if and only if $I(V)$ is a prime ideal.

Recall that $I(V)$ is prime when $f_1 f_2 \in I(V)$ implies $f_1 \in I(V)$ or $f_2 \in I(V)$. Geometrically, the ideal is not prime when we can find two functions where the product is zero on V but are individually not zero on all of V.

Proof. Recall that $I(V_1 \cup V_2) = I(V_1) \cap I(V_2)$. Suppose V were reducible, so $V = V_1 \cup V_2$ where $V_1, V_2 \neq V$. In particular, $V_1 \not\subseteq V_2 \not\subseteq V_1$. Now, let $I_j = I(V_j)$, giving $I_1 \not\supseteq I_2 \not\supseteq I_1$, and $I(V) = I_1 \cap I_2$. Therefore, there exists $f_1 \in I_1 \setminus I_2$ and $f_2 \in I_2 \setminus I_1$. Each f_i is not an element of $I(V)$, but $f_1 f_2 \in I(V)$. So $I(V)$ cannot be prime.

Conversely, suppose $I(V)$ is not prime, so $f_1 f_2 \in I(V)$ but $f_1, f_2 \notin I(V)$. Define $V_1 = V \cap \mathbb{V}(f_1)$ and $V_2 = V \cap \mathbb{V}(f_2)$. Since neither f_i is contained in $I(V)$, $V_i \neq V$. Also, if $P \in V$, we have $f_1(P) f_2(P) = 0$, so $P \in V_1 \cup V_2$. So V is reducible. □

Example. Let $V = \mathbb{V}(XY) \subset \mathbb{A}^2$. Then $V = \mathbb{V}(X) \cup \mathbb{V}(Y)$ is a decomposition of V into irreducible components. Indeed, $\mathbb{V}(X)$ is irreducible, as $I(\mathbb{V}(X)) = (X)$ is a prime ideal in $\mathbb{C}[X, Y]$, and similarly for Y.

2. Structures on varieties

2.1. Coordinate rings

Consider a polynomial $f \in \mathbb{C}[\mathbf{X}]$. We obtain a function $f : \mathbb{A}^n \to \mathbb{A}^1$, If $V \subseteq \mathbb{A}^n$ and $f, g \in \mathbb{C}[\mathbf{X}]$, we are interested in when f, g induce the same set-theoretic function on V. We intend to show that f, g induce the same function if and only if $f - g \in I(V)$. Therefore, we can study polynomials modulo this relation by taking the quotient with respect to this ideal.

Definition. Let $V \subseteq \mathbb{A}^n$ be a variety. The *coordinate ring* of V, or the *ring of regular functions* of V, is defined as ${}^{\mathbb{C}[\mathbf{X}]}\!/_{I(V)}$, denoted $\mathbb{C}[V]$ or $\mathcal{O}(V)$.

Corollary. Let V be a variety. Then V is irreducible if and only if $\mathbb{C}[V]$ is an integral domain.

Remark. $\mathbb{C}[V]$ does not precisely determine V or $I(V)$. For instance, consider a surjective homomorphism $\theta : \mathbb{C}[\mathbf{X}] \to \mathbb{C}[V]$, then $\ker\theta = I$ is an ideal, and $\mathbb{V}(I)$ is a variety with coordinate ring $\mathbb{C}[V]$. However, there is not a unique such homomorphism in general. For instance, $\mathbb{C}[X] \simeq {}^{\mathbb{C}[X, Y]}\!/_{(Y)}$.

Definition. Let $I \lhd \mathbb{C}[\mathbf{X}]$. We define the *radical ideal* of I to be

$$\sqrt{I} = \{f \in \mathbb{C}[\mathbf{X}] \mid \exists m > 0,\ f^m \in I\}$$

This is an ideal. $\sqrt{\sqrt{I}} = \sqrt{I}$. Note that $\mathbb{V}(I) = \mathbb{V}\left(\sqrt{I}\right)$.

Theorem (strong form of Hilbert's Nullstellensatz)**.** Let $I \lhd \mathbb{C}[\mathbf{X}]$ be an ideal, and $V = \mathbb{V}(I)$. Then $I(V) = \sqrt{I}$.

Therefore, the map $V \mapsto I(V)$ maps precisely onto the space of radical ideals, ideals which are equal to their radicals.

Example. Let $V = \{0\} \in \mathbb{A}^1$. We can write $V = \mathbb{V}(X^2)$, so its coordinate ring is

$${}^{\mathbb{C}[X]}\!/_{I(\mathbb{V}(X^2))} = {}^{\mathbb{C}[X]}\!/_{\sqrt{(X^2)}} = {}^{\mathbb{C}[X]}\!/_{(X)} \simeq \mathbb{C}$$

In building the coordinate ring, we forget the structure of X^2. If we had instead considered ${}^{\mathbb{C}[X]}\!/_{(X^2)}$, we would have certain nonzero elements whose squares are zero.

2.2. Morphisms

Let $V \subseteq \mathbb{A}^n$ and $W \subseteq \mathbb{A}^m$ be affine varieties.

Definition. A *regular map* or *morphism* from V to W is a function $\varphi : V \to W$ such that there exist elements $f_1, \dots, f_m \in \mathbb{C}[V]$ such that

$$\varphi(P) = (f_1(P), \dots, f_m(P))$$

for all $P \in V$.

The set of all morphisms from V to W is denoted $\mathrm{Mor}(V, W)$.

Example. The morphisms V to $\mathbb{A}^1$ are precisely the functions in the coordinate ring $\mathbb{C}[V]$.

Example. Linear projections $\mathbb{A}^n \to \mathbb{A}^m$ are morphisms. More generally, linear transformations and affine translations are also morphisms.

Example. If $V \subseteq W \subseteq \mathbb{A}^n$ where V, W are varieties, then the inclusion map $V \hookrightarrow W$ is a morphism.

Proposition. Let $\varphi : V \to W, \psi : W \to Z$ be morphisms. Then the composite map $\psi \circ \varphi$ is a morphism $V \to Z$.

Proof. The composition of polynomials is a polynomial. □

2.3. Pullbacks

Definition. Let $\varphi : V \to W$ be a morphism, and let $g \in \mathbb{C}[W]$. Then, the *pullback* is $\varphi^\star(g) = g \circ \varphi : V \to \mathbb{C}$. Note that $\varphi^\star(g) \in \mathbb{C}[V]$, so $\varphi^\star$ gives a map $\mathbb{C}[W] \to \mathbb{C}[V]$.

Remark. This map $\varphi^\star$ is a ring homomorphism, and restricts to the identity on $\mathbb{C}$.

Definition. A ring homomorphism $\mathbb{C}[X] \to \mathbb{C}[Y]$ that restricts to the identity on $\mathbb{C}$ is called a $\mathbb{C}$-*algebra homomorphism*.

Theorem. Let $V \subseteq \mathbb{A}^n, W \subseteq \mathbb{A}^m$ be affine varieties. The map $\alpha : \varphi \mapsto \varphi^\star$ defines a bijection from $\mathrm{Mor}(V, W)$ to the space of $\mathbb{C}$-algebra homomorphisms $\mathbb{C}[W] \to \mathbb{C}[V]$.

Proof. Let $y_1, \dots, y_n \in \mathbb{C}[W]$ be the coordinate functions on W, which are the restrictions of the standard linear coordinate functions on $\mathbb{A}^n$.

First, we show injectivity of α. Let $\varphi : V \to W$ be a morphism. For any point $P \in V$,

$$\varphi(P) = (y_1(\varphi(P)), \dots, y_m(\varphi(P))) = (\varphi^\star(y_1)(P), \dots, \varphi^\star(y_n)(P))$$

So φ is determined by the values of $\varphi^\star(y_1), \dots, \varphi^\star(y_n)$.

Now we show its surjectivity. Let $\lambda : \mathbb{C}[W] \to \mathbb{C}[V]$ be a $\mathbb{C}$-algebra homomorphism, and let $f_i = \lambda(y_i) \in \mathbb{C}[V]$. We can now define the map $\varphi = (f_1, \dots, f_m) : V \to \mathbb{A}^m$. We will show that φ has image contained in W, so that we have $\varphi : V \to W$, which then shows that φ is a morphism $V \to W$. For $P \in V$, we must show $g(\varphi(P)) = 0$ for all $g \in I(W)$. We obtain $g(f_1(P), \dots, f_m(P)) = \lambda(g)(P)$. But $g = 0$ in $\mathbb{C}[W]$, so $g(\varphi(P)) = 0$ as required. Hence $\varphi : V \to W$ is a morphism, and $\lambda = \varphi^\star$ since $\varphi^\star(y_i) = f_i = \lambda(y_i)$. □

Definition. Two affine varieties V, W are *isomorphic* if we have $\varphi : V \to W, \psi : W \to V$ where $\varphi \circ \psi = \mathrm{id}_W$ and $\psi \circ \varphi = \mathrm{id}_V$.

Theorem. V is isomorphic to W if and only if $\mathbb{C}[V]$ is isomorphic to $\mathbb{C}[W]$ as $\mathbb{C}$-algebras.

Proof. Use the previous theorem. □

Example. The affine line $\mathbb{A}^1$ is isomorphic to the twisted cubic $\{(t, t^2, t^3) \mid t \in \mathbb{C}\}$. This can be easily shown by calculating the coordinate rings explicitly.

Example. Let $V \subseteq \mathbb{A}^2$ be given by $X_1X_2(X_1 - X_2) = 0$. This is the union of three lines, intersecting at the origin. Let $W \subseteq \mathbb{A}^3$ be given by $X_1X_2 = X_2X_3 = X_3X_1 = 0$, which is also a union of three lines, which in this case are the coordinate axes. These are not isomorphic as varieties, because their coordinate rings are not isomorphic, which can be easily shown using tangent spaces, defined in later sections. Note, however, that V and W are homeomorphic in the Euclidean topology.

2.4. Rational functions

Definition. Let $V \subseteq \mathbb{A}^n$ be an irreducible variety. Its *function field*, *field of rational functions*, or *field of meromorphic functions* is the field of fractions $\mathbb{C}(V) = FF(\mathbb{C}[V])$ of $\mathbb{C}[V]$.

Remark. Since V is irreducible, $I(V)$ is prime, so $\mathbb{C}[V]$ is an integral domain. This allows us to construct the field of fractions.

Definition. Let φ be a rational function. A point $P \in V$ is called *regular* if φ can be expressed as a ratio $\frac{f}{g}$ with $g(P) \neq 0$.

Remark. If $\varphi = \frac{f}{g}$, we obtain a well-defined function $\varphi : V \setminus \mathbb{V}(g) \to \mathbb{C}$. The domain is an open set in V, since $\mathbb{V}(g)$ is Zariski closed.

Example. Consider the rational function $X_1^2/X_2 \in \mathbb{C}(\mathbb{A}^2)$. This defines a map on the complement of the X_2-axis. Note that X^3/X_1X_2 defines the same function, but only on points other than $\mathbb{V}(X_1X_2)$. Note that $X^3/X_1X_2 = X_1^2/X_2 \in \mathbb{C}(\mathbb{A}^2)$, so we cannot quite think of elements of $\mathbb{C}(\mathbb{A}^2)$ as functions.

Remark. A rational function on V can be thought of as a pair (U, f) with $U \subseteq V$ Zariski open, such that f is a function $U \to \mathbb{C}$. We define the equivalence relation $(U, f) \sim (U', f')$ if f, f' agree on some nonempty Zariski open set contained in U and U'. Note that if V is irreducible, every nonempty open subset is dense in the Zariski topology.

Definition. A *local ring* is a ring R that contains a unique maximal ideal.

Definition. Let V be an irreducible variety, and let P. The *local ring of V at P* is $\mathcal{O}_{V,P} = \{f \in \mathbb{C}(V) \mid f \text{ is regular at } P\}$.

Proposition. The local ring of an irreducible variety V at a point P is a local ring. Its unique maximal ideal is

$$\mathfrak{m}_{V,P} = \{f \in \mathcal{O}_{V,P} \mid f(P) = 0\} = \ker(f \mapsto f(P))$$

Further, the invertible elements of $\mathcal{O}_{V,P}$ are precisely those f such that $f(P) \neq 0$.

The proof follows from the following more general lemma.

Lemma. A ring R is a local ring if and only if $R \setminus R^\star$ is an ideal. If so, the unique maximal ideal is $R \setminus R^\star$.

Proof. If $A \trianglelefteq R$ is an ideal, then $A \neq R$ if and only if $A \subseteq R \setminus R^\star$, because if any unit lies in A, it must be all of R. Hence, if $R \setminus R^\star$ is an ideal, it is automatically the unique maximal ideal.

Conversely, let R be a local ring with unique maximal ideal $\mathfrak{m}$. Then $\mathfrak{m} \subseteq R \setminus R^\star$, and if $x \in R \setminus R^\star$ we must have $(x) \neq R$, so $(x) \subseteq \mathfrak{m}$ by maximality. Hence $\mathfrak{m} = R \setminus R^\star$. □

Note that this proves the previous proposition, as $\frac{f}{g} \in \mathcal{O}_{V,P}$ is invertible if and only if $\left(\frac{f}{g}\right)(P) \neq 0$.

Example. Let

$$R = \left\{ \frac{f}{g} \in \mathbb{C}(t) \,\middle|\, \text{ignoring factors, g(0)} \neq 0 \right\} = \mathcal{O}_{\mathbb{A}^1,0}$$

Here, the maximal ideal is (t), and $R/(t) = \mathbb{C}$.

Let $S = \mathbb{C}[\![t]\!]$ be the ring of formal power series in t. This is a local ring by the lemma; its maximal ideal is (t). Note that in fact $R \subseteq S$.

3. Projective varieties

We will construct the projective space $\mathbb{P}^n$, which will be an upgrade to $\mathbb{A}^n$; it is not immediately obvious why $\mathbb{P}^n$ is considered 'better'. Projective space has some interesting properties, such as:

- every pair of lines in $\mathbb{P}^2$ that are distinct meet at a unique point;
- if V is a projective variety (defined shortly) in $\mathbb{P}^2$ defined by a degree d polynomial, if V is a manifold then V is homeomorphic in the Euclidean topology to a closed orientable topological surface of genus $\binom{d-1}{2}$.
- $\mathbb{P}^n$ is compact in the Euclidean topology, but $\mathbb{A}^n$ is not.

3.1. Definition

Definition. Let U be a finite-dimensional complex vector space. The *projectivisation* of U, written $\mathbb{P}(U)$, is the set of lines in U through the origin $\mathbf{0} \in U$. Define $\mathbb{P}^n = \mathbb{P}(\mathbb{C}^{n+1})$.

We usually index the coordinates on $\mathbb{C}^{n+1}$ with indices $0, \dots, n$. A line in $\mathbb{C}^{n+1}$ is therefore given by $\{(a_0 t, \dots, a_n t) \mid t \in \mathbb{C}\}$, and is written $L_{(a_0,\dots,a_n)}$, where not all a_i are zero. We write $(a_0 : a_1 : \dots : a_n)$ for the corresponding element of $\mathbb{P}^n$. Therefore,

$$\mathbb{P}^n = {\{(a_0, \dots, a_n) \mid a_i \in \mathbb{C}, \text{not all } a_i = 0\}}\big/_{\text{scaling by } \mathbb{C}^\star}$$

For example, $(2 : 1 : -2) = (4 : 2 : -4) \in \mathbb{P}^2$.

We can decompose $\mathbb{P}^1$ as

$$\begin{aligned}\{(a_0 : a_1) \mid a_0 \neq 0\} \cup \{(a_0 : a_1) \mid a_0 = 0\} &= \{(1 : z) \mid z \in \mathbb{C}\} \cup \{(0 : 1)\} \\ &= \mathbb{A}^1 \cup \text{a point at infinity}\end{aligned}$$

More generally,

$$\mathbb{P}^n = \{(a_0 : \dots : a_n) \mid a_0 \neq 0\} \cup \{(0 : a_1 : \dots : a_n)\} = \mathbb{A}^n \amalg \mathbb{P}^{n-1}$$

By induction, $\mathbb{P}^n = \mathbb{A}^n \cup \mathbb{A}^{n-1} \cup \dots \cup \mathbb{A}^1 \cup$ a point, where the terms other than $\mathbb{A}^n$ are considered 'at infinity'.

Definition. The *Zariski* (respectively *Euclidean*) topology on projective space is the quotient topology for the subspace topology for the Zariski (respectively Euclidean) topology on $\mathbb{C}^{n+1} \setminus \{\mathbf{0}\}$, where $\mathbb{P}^n = {\mathbb{C}^{n+1} \setminus \{\mathbf{0}\}}\big/_{\sim}$ and $\mathbb{C}^{n+1} \setminus \{0\} \subseteq \mathbb{C}^{n+1}$.

There is a copy of S^{2n+1} inside $\mathbb{C}^{n+1} \setminus \{\mathbf{0}\}$, which therefore surjects onto $\mathbb{P}^n$.

Corollary. $\mathbb{P}^n$ is compact.

Proof. It is the continuous image of the compact set S^{2n+1}. □

Definition. For $0 \leq j \leq n$, we define the *jth coordinate hyperplane* is the set $H_j = \{(\mathbf{a}_i) \mid a_j = 0\} \subseteq \mathbb{P}^n$.

We can naturally identify H_j with $\mathbb{P}^{n-1}$.

Definition. The *jth standard affine patch* U_j is the complement of H_j.

There is a natural bijection $U_j \to \mathbb{A}^n$ by mapping $(a_0 : \dots : a_n)$ to $\left(\frac{a_0}{a_j}, \dots, \frac{\widehat{a_j}}{a_j}, \dots, \frac{a_n}{a_j}\right)$ where the hat denotes 'forgetting' that element of the tuple. The inverse function maps $(b_1, \dots, b_n)$ to $(b_1 : \dots : b_{j-1} : 1 : b_j : \dots : b_n)$. We observe that $\{U_j\}_{j=0}^n$ is an open cover of $\mathbb{P}^n$ in the Zariski topology.

3.2. Projective varieties

Example. Consider the polynomial $X_0 + 1 \in \mathbb{C}[X_0, X_1]$. Note that $X_0 + 1$ does not define a function on $\mathbb{P}^1$. For example, $(-1 : 0) = (1 : 0)$, but $X_0 + 1$ vanishes on the first representative and not the second. The vanishing locus of $X_0 + 1$ on $\mathbb{P}^1$ is therefore undefined. Therefore, we need a slightly more subtle definition of a variety in projective space.

Definition. A *monomial* in $\mathbb{C}[\mathbf{X}] = \mathbb{C}[X_0, \dots, X_n]$ is an element of the form $X_0^{d_0} X_1^{d_1} \dots X_n^{d_n}$ where $d_i \geq 0$. A *term* is a nonzero multiple of a monomial. The *degree* of a term $cX_0^{d_0} \dots X_n^{d_n}$ is $\sum_{i=0}^n d_i$. A *homogeneous polynomial* of degree d is a finite sum of terms of degree d.

Any polynomial can be uniquely decomposed as a sum of homogeneous polynomials of different degree; we write $f = \sum_{i=0}^\infty f_{[i]}$ where the $f_{[i]}$ are homogeneous of degree i. Note that this sum is always finite.

Lemma. Let $f \in \mathbb{C}[\mathbf{X}]$ be homogeneous, and let $(a_0, \dots, a_n) \in \mathbb{C}^{n+1} \setminus \{\mathbf{0}\}$. Then, if $f(\mathbf{a}) = 0$, we have $f(\lambda \mathbf{a}) = 0$ for all $\lambda \in \mathbb{C}^\star$.

Proof. Trivial by checking the definitions. □

Corollary. Let $f \in \mathbb{C}[\mathbf{X}]$ be homogeneous. Then

$$\mathbb{V}(f) = \{P \in \mathbb{P}^n \mid f(\mathbf{a}) = 0 \text{ for any (or every) representative of } P\}$$

is well-defined.

Definition. An ideal $I \trianglelefteq \mathbb{C}[\mathbf{X}]$ is called *homogeneous* if it can be generated by homogeneous polynomials (of potentially different degrees).

Lemma. Let $I \trianglelefteq \mathbb{C}[\mathbf{X}]$ be an ideal. Then I is homogeneous if and only if whenever $f \in I$, all of the homogeneous parts $f_{[r]}$ are also contained in I.

Proof. Suppose I is homogeneous. Then let g_j be homogeneous generators of I of degree d_j. Writing $f = \sum h_j g_j$ for arbitrary $h_j \in \mathbb{C}[\mathbf{X}]$, we can split each h_j into its pieces $h_{j[r]}$.

Now, $h_{j[r]}g_j \in I$ is homogeneous, and its degree is rd_j. Hence, $f_{[r]} = \sum_j h_{j[r-dj]}g_j \in I$ as required. The converse is trivial by decomposing the generators of the ideal. □

Definition. Let $I \trianglelefteq \mathbb{C}[\mathbf{X}]$ be a homogeneous ideal. Then, the *vanishing locus* is $\mathbb{V}(I) = \{P = (\mathbf{a}_i) \in \mathbb{P}^n \mid \forall f \in I,\ f((\mathbf{a}_i)) = 0\}$. A *projective variety* in $\mathbb{P}^n$ is any set of this form.

Note that we could have defined the vanishing locus using the quantifier 'for all *homogeneous* $f \in I$'.

Example. Let $U \subseteq \mathbb{C}^{n+1}$ be any vector subspace. Let the projectivisation of U is a subset of $\mathbb{P}^n$, and is a projective variety. More concretely, $U = \left\{\mathbf{v} \in \mathbb{C}^{n+1} \mid \forall j,\ \sum_{i=0}^n a_i^{(j)} v_i = 0\right\}$ for a subset $\mathbf{a}^{(j)} = (a_0^{(j)}, \dots, a_n^{(j)})$, as a vector subspace is the kernel of some linear map. Therefore, $\mathbb{P}(U) = \mathbb{V}(I)$ where I is the ideal generated by $F_j = \sum_i a_i^{(j)} X_i \in \mathbb{C}[\mathbf{X}]$. More generally, a projective linear space is the projectivisation of a linear subspace. Hence, projective linear spaces in $\mathbb{P}^n$ are in bijection with linear subspaces in $\mathbb{C}^{n+1}$.

$GL_{n+1}(\mathbb{C})$ acts on $\mathbb{P}^n$ coordinatewise. The normal subgroup of scalar matrices $\mathbb{C}^\star \subseteq GL_{n+1}(\mathbb{C})$ acts trivially on $\mathbb{P}^n$. The quotient is written $PGL_n(\mathbb{C}) = {}^{GL_{n+1}(\mathbb{C})}\!/_{\mathbb{C}^\star}$, and acts transitively on $\mathbb{P}^n$.

Example. The *Segre surface* is the hypersurface $S_{11} = \mathbb{V}(X_0X_3 - X_1X_2) \subseteq \mathbb{P}^3$. Consider the map $\sigma_{11} : \mathbb{P}^1 \times \mathbb{P}^1 \to \mathbb{P}^3$ given by $\sigma_{11}((a_0 : a_1), (b_0 : b_1)) = (a_0b_0 : a_0b_1 : a_1b_0 : a_1b_1)$. One can show that this map is well-defined, and in fact, $\operatorname{Im} \sigma_{11} = S_{11}$.

First, consider the map $\mathbb{C}^2 \times \mathbb{C}^2 \to \mathbb{C}^4$ where we identify $\mathbb{C}^4$ with the space of 2×2 matrices on $\mathbb{C}$, given by the outer product. More precisely, $(v, w) \mapsto vw^\intercal$. The image of this map is precisely the set of matrices of rank at most 1. Hence, the image is the vanishing locus of $X_0X_3 - X_1X_2$, the determinant of such a matrix.

3.3. Homogenisation and projective closure

Recall that $\mathbb{P}^n = U_0 \cup \dots \cup U_n$, where $U_i = \mathbb{P}^n \setminus \mathbb{V}(X_i)$. We therefore have the following different descriptions of a Zariski topology on $\mathbb{P}^n$:

(i) the quotient of the subspace of the Zariski topology on $\mathbb{C}^{n+1}$;

(ii) define that V is Zariski-closed if and only if $V = \mathbb{V}(I)$ where $I \lhd \mathbb{C}[\mathbf{X}]$ is homogeneous;

(iii) the gluing topology: define that a subset $Z \subseteq \mathbb{P}^n$ is closed if $Z \cap U_i$ is closed for all i, as the U_i are isomorphic to $\mathbb{A}^n$.

These three constructions coincide.

If $V \subseteq \mathbb{P}^n$ is a projective variety, consider $U_0 \cap V \subseteq U_0$. If $V = \mathbb{V}(I)$, then $U_0 \cap V = \mathbb{V}(I_0)$ where $I_0 = \{f = F(1, Y_1, \dots, Y_n) \mid F \in I \text{ homogeneous}\} \subseteq \mathbb{C}[Y_1, \dots, Y_n]$. Identifying U_0 with $\mathbb{A}^n$ with coordinates $Y_1, \dots, Y_n$ (so $Y_j = \frac{X_j}{X_0}$), $V \cap U_0$ is naturally an affine variety.

Conversely, let $W \subseteq \mathbb{A}^n$ be an affine variety, and identify $\mathbb{A}^n$ with U_0. Then, the Zariski closure $\overline{W}$ of W inside $\mathbb{P}^n$ is a projective variety. We are interested in studying the precise projective varieties obtained in this way.

Definition. Let $f \in \mathbb{C}[Y_1, \dots, Y_n]$ be an arbitrary polynomial of total degree d. The *homogenisation* of f, written F or f^h, is

$$f^h(X_0, \dots, X_n) = X_0^d f\left(\frac{X_1}{X_0}, \dots, \frac{X_n}{X_0}\right) \in \mathbb{C}[X_0, \dots, X_n]$$

This is homogeneous of degree d. Similarly, if I is an ideal in $\mathbb{C}[Y_1, \dots, Y_n]$, its homogenisation $I^\star = I^h$ is the ideal generated by the homogenisation of the elements $f \in I$; this is a homogeneous ideal in $\mathbb{C}[X_0, \dots, X_n]$. Given an affine variety $V \subseteq \mathbb{A}^n$, the *projective closure* of V is $\mathbb{V}(I(V)^h) \subseteq \mathbb{P}^n$.

Example. Let $f(Y_1, Y_2) = 1 + Y_1^2 + Y_1 Y_2^2$. Its homogenisation is $f^h(X_0, X_1, X_2) = X_0^3 + X_0 X_1^2 + X_1 X_2^2$.

Remark. Let $I = (f_1, \dots, f_r) \subseteq \mathbb{C}[Y_1, \dots, Y_n]$, and let $J = (f_1^h, \dots, f_r^h)$. Typically, $J \neq I^h$. If I is principal, this holds: $I = (f)$ implies $I^h = (f^h)$.

Proposition. Let $V \subseteq \mathbb{A}^n$ be an affine variety. Then, the Zariski closure $\overline{V} \subseteq \mathbb{P}^n$ given by identifying $U_0 = \mathbb{A}^n$ coincides with the projective closure $\mathbb{V}(I(V)^h) \subseteq \mathbb{P}^n$.

Proof. Let I be an ideal in $\mathbb{C}[Y_1, \dots, Y_n]$, and let $V = \mathbb{V}(I)$. Let $\overline{V}$ be the Zariski closure. Let I^h be the homogenisation of the ideal. Then $\mathbb{V}(I^h)$ is Zariski closed, and contains V. We will show that this is the smallest such set.

Suppose $Y \supseteq V$ is closed, so $Y = \mathbb{V}(I')$ where I' is homogeneous. Any homogeneous element in I' can be written as $X_0^d f^h$ for some $f \in \mathbb{C}[Y_1, \dots, Y_n]$. Now, $X_0^d f^h = 0$ on $V \subseteq \mathbb{P}^n$, so $f = 0$ on $V \subseteq \mathbb{A}^n$. Hence $f \in I(V) = \sqrt{I}$ by the Nullstellensatz. So $f^m \in I$ for some $m > 0$, so $(f^m)^h = (f^h)^m \in I^h$. Hence $f^h \in \sqrt{I^h}$, so $X_0^d f^h \in \sqrt{I^h}$. Therefore, $I' \subseteq \sqrt{I^h}$. □

Remark. Let $V \subseteq \mathbb{P}^n$, and let $W = V \cap U_0 \subseteq \mathbb{A}^n$. Then $\overline{W} \subseteq \mathbb{P}^n$ is not in general equal to V. For example, let $V = \mathbb{V}(X_0)$, so $W = \varnothing$ and $\overline{W} = \varnothing$. This ambiguity arises due to the X_0^d term required in the above proof when dehomogenising a polynomial.

This shows that the topological notion of the Zariski closure and the algebraic notion of the projective closure agree.

Example. Let $V \subseteq \mathbb{P}^2$ be given by $\mathbb{V}(X_0 X_1 - X_2^2)$. We obtain $V_0 \subseteq U_0$ given by setting $X_0 = 1$, $V_1 \subseteq U_1$ given by setting $X_1 = 1$, and $V_2 \subseteq U_2$ given by setting $X_2 = 1$. We find $V_0 = \mathbb{V}(Y_1 - Y_2^2)$ which is a parabola, and V_1 is similar. $V_2 = \mathbb{V}(X_0 X_1 - 1)$, which is a rectangular hyperbola.

Theorem. Let $Q \subseteq \mathbb{P}^n$ be given by $\mathbb{V}(f)$ where f is a homogeneous quadratic polynomial. Then, after a change of coordinates $A \in PGL_n(\mathbb{C})$, Q has the form $\mathbb{V}(X_0^2 + \dots + X_r^2)$ where r is the rank of the quadratic form f.

Proof. Use the results from IB Linear Algebra. □

Theorem (projective Nullstellensatz)**.** If $\mathbb{V}(I) = \varnothing \subseteq \mathbb{P}^n$ where I is a homogeneous ideal, then $I \supseteq (X_0^m, \dots, X_n^m)$ for some $m \in \mathbb{N}$. Further, if $V = \mathbb{V}(I) \neq \varnothing$, then $I^h(V) = \sqrt{I}$, where $I^h(V)$ is the ideal generated by homogeneous polynomials vanishing on V.

Proof. We reduce to the affine case. Let I be a homogeneous ideal, and let $V^a = \mathbb{V}(I) \subseteq \mathbb{A}^{n+1}$. Note that $\mathbf{0} \in V^a$, assuming $V \neq \varnothing$. Then there is a continuous map $V^a \setminus \{0\} \to V$ obtained by the restriction of $\mathbb{A}^{n+1} \setminus \{\mathbf{0}\} \to \mathbb{P}^n$. Moreover, this map is surjective, so is a quotient map. Note that V is empty if and only if $V^a = \{\mathbf{0}\}$. So the result holds by the affine Nullstellensatz. The second part is similar. □

Let V be a projective variety in $\mathbb{P}^n$. If $W \subseteq V$ is a variety closed in V, we say W is a *closed subvariety* of V. The complement $V \setminus W$ is an *open subvariety*. The closed (respectively open) subvarieties of V satisfy the axioms of the closed (open) sets of a topology. We say V is irreducible if V cannot be written as $V_1 \cup V_2$ for proper closed subvarieties V_1, V_2.

Proposition. (i) Every projective variety is a finite union of irreducible varieties.

(ii) V is irreducible if and only if $I^h(V)$ is prime.

Proof. Part (i) is identical to the affine case. For part (ii), first observe that if I is a homogeneous ideal which is not prime, we can find homogeneous $F, G \notin I$ such that $FG \in I$, as I is generated by homogeneous elements. Then the proof for the affine case works as before. □

Let $S \subseteq V$ be a subset. S is Zariski dense in V if and only if every homogeneous polynomial that vanishes on S vanishes on V.

Proposition. Let $V \subseteq \mathbb{P}^n$ be an irreducible projective variety. Let $W \subsetneq V$ be a proper closed subvariety. Then, $V \setminus W$ is dense in V.

Intuitively, W is lower-dimensional than V, and V with a lower-dimensional set removed is dense.

Proof. Let $f \in \mathbb{C}[\mathbf{X}]$ be a homogeneous polynomial that vanishes on $V \setminus W$. As $W \neq V$, there exists a polynomial $g \in I^h(W) \setminus I^h(V)$ by the projective Nullstellensatz. Then, fg vanishes on all of V. But $I^h(V)$ is prime as V is irreducible, so $f \in I^h(V)$. □

3.4. Rational functions

Homogeneous polynomials have well-defined zero sets in $\mathbb{P}^n$, but not a well-defined value. Therefore, we cannot define a coordinate ring $\mathbb{C}[V]$ in an analogous way. However, a ratio of homogeneous polynomials of the same degree does have a well-defined value on $\mathbb{P}^n$ away from the vanishing locus of the denominator.

Definition. Let $V \subseteq \mathbb{P}^n$ be an irreducible projective variety. The *function field* or *field of rational functions* is

$$\mathbb{C}(V) = \left\{ \frac{F}{G} \,\middle|\, F, G \in \mathbb{C}[\mathbf{X}] \text{ homogeneous and have the same degree, } G \notin I^h(V) \right\} \Big/ \sim$$

where $\frac{F_1}{G_1} \sim \frac{F_2}{G_2}$ if $F_1 G_2 - F_2 G_1 \in I^h(V)$.

Lemma. The relation $\sim$ is an equivalence relation.

Proof. Reflexivity and symmetry are clear. Now suppose that $\frac{F_1}{G_1} \sim \frac{F_2}{G_2} \sim \frac{F_3}{G_3}$, so $F_2G_1 - F_1G_2 \in I^h(V)$ and $F_3G_2 - F_2G_3 \in I^h(V)$. Consider $F_1G_3 - F_3G_1$. Multiplying by G_2, $F_1G_2G_3 - F_3G_1G_2$. Since $G_2 \notin I^h(V)$, primality of $I^h(V)$ implies that it suffices to show $F_1G_2G_3 - F_3G_1G_2 \in I^h(V)$. In the ring $\mathbb{C}[\mathbf{X}] \big/ I^h(V)$, we have relations $F_1G_2 = F_2G_1$ and $F_3G_2 = F_2G_3$. Hence $F_1G_2G_3 - F_3G_1G_2 = 0$ in $\mathbb{C}[\mathbf{X}] \big/ I^h(V)$. □

Note that $\mathbb{C}(V)$ is a field.

Proposition. The field $\mathbb{C}(V)$ is a finitely generated field extension of $\mathbb{C}$.

Note that $\mathbb{C}(t)$ is finitely generated as a field, but not finitely generated as a $\mathbb{C}$-module or a $\mathbb{C}$-algebra.

Proof. Suppose $V \neq \varnothing$. Then, there is some coordinate function X_i that is nonzero on V; without loss of generality let $i = 0$. We claim that $\frac{X_1}{X_0}, \dots, \frac{X_n}{X_0}$ generate $\mathbb{C}(V)$ over $\mathbb{C}$. Explicitly, if $\frac{F}{G}$ is a degree 0 ratio, it can be written in terms of the $\frac{X_j}{X_0}$ and the field operations. It suffices to show the result holds when $\frac{F}{G}$ is of the form $\frac{M}{G}$ where M is a monomial. Then, it suffices to show the result for $\frac{G}{M}$ where M is a monomial by taking reciprocals. Hence, it suffices to show the result for $\frac{M}{M'}$ where M, M' are monomials, and this is trivial. □

Corollary. Let $V \subseteq \mathbb{P}^n$ be an irreducible projective variety, not contained in the hyperplane $\{X_0 = 0\}$. Let $V_0 = V \cap U_0$, where $U_0 \simeq \mathbb{A}^n$ is the first affine patch. Then, $\mathbb{C}(V_0) = \mathbb{C}(V)$, where $\mathbb{C}(V_0) = FF(\mathbb{C}[V_0])$.

Proof. V_0 has coordinate ring

$$\mathbb{C}\left[\frac{X_1}{X_0}, \dots, \frac{X_n}{X_0}\right] \Big/ I(V_0)$$

Hence, $\mathbb{C}(V_0) = FF(\mathbb{C}[V_0])$ is generated by the $\frac{X_j}{X_0}$. □

Definition. Let $\varphi \in \mathbb{C}(V)$ and let $P \in V$. We say that φ is *regular* or *defined* at P if φ can be expressed as $\frac{F}{G}$ where F, G are homogeneous of the same degree with $G(P) \neq 0$. There is a partial function from the set of regular points of V to $\mathbb{C}$.

Definition. The *local ring* of V at P, written $\mathcal{O}_{V,P}$, is the set of $\varphi \in \mathbb{C}(V)$ such that φ is regular at P. This is a subring of $\mathbb{C}(V)$, which is a local ring in the sense of commutative algebra.

Proposition. Let $V \subseteq \mathbb{P}^n$ be an irreducible projective variety not contained in $\{X_0 = 0\}$. Let $V_0 = V \cap U_0$ where $U_0 = \{X_0 = 0\}$. Let P be a point in V_0. Then, there is a natural isomorphism $\mathcal{O}_{V,P} \to \mathcal{O}_{V_0,P}$ respecting the isomorphism $\mathbb{C}(V) \simeq \mathbb{C}(V_0)$.

Proof. Follows by unfolding the definitions. □

3.5. Rational maps

Let $F_0, \dots, F_m \in \mathbb{C}[\mathbf{X}] = \mathbb{C}[X_0, \dots, X_n]$ be homogeneous of the same degree d. Define $\mathbf{F} = (F_0, \dots, F_m) \colon \mathbb{C}^{n+1} \to \mathbb{C}^{m+1}$.

Proposition. The map $\mathbf{F}$ descends to a well-defined map of sets $\varphi \colon \mathbb{P}^n \setminus \bigcap_j \mathbb{V}(F_j) \to \mathbb{P}^m$. If P is represented by $\mathbf{a} = (a_0, \dots, a_n)$, then $\varphi(P)$ is represented by $(F_0(\mathbf{a}), \dots, F_m(\mathbf{a}))$.

Proof. Since all F_j are homogeneous of the same degree d, $\lambda\mathbf{a} = (\lambda a_0, \dots, \lambda a_n)$ gives

$$(F_0(\lambda\mathbf{a}), \dots, F_m(\lambda\mathbf{a})) = \lambda^d(F_0(\mathbf{a}), \dots, F_m(\mathbf{a}))$$

which is equivalent to $\varphi(P)$. □

We will denote such maps $\mathbf{F} = (F_0, \dots, F_m)$ by $\varphi \colon \mathbb{P}^n \dashrightarrow \mathbb{P}^m$.

Let G be a nonzero homogeneous polynomial in $X_0, \dots, X_n$. Given $\mathbf{F} \colon \mathbb{P}^n \dashrightarrow \mathbb{P}^m$, we can also consider $G\mathbf{F} = (GF_0, \dots, GF_n) \colon \mathbb{P}^n \dashrightarrow \mathbb{P}^m$. Observe that the maps $\mathbf{F}$ and $G\mathbf{F}$ have different domains, but coincide at points where they are both defined. Note that there is a 'best' representative $\mathbf{F}$, as $\mathbb{C}[\mathbf{X}]$ is a unique factorisation domain, but we will not use this notion here, because not all rings that we will use are unique factorisation domains.

Definition. Let $V \subseteq \mathbb{P}^n$ be an irreducible projective variety. Let $F_0, \dots, F_m$ be homogeneous polynomials in $\mathbb{C}[X_0, \dots, X_n]$ of fixed degree d, and not all contained in $I^h(V)$. They determine a map of sets $V \setminus \bigcap_j \mathbb{V}(F_j) \to \mathbb{P}^n$ by the previous construction. Two such tuples $(F_0, \dots, F_m)$ and $(G_0, \dots, G_m)$ are said to *determine the same map* if $F_iG_j - F_jG_i \in I^h(V)$. A *rational map* from V to $\mathbb{P}^m$ is an equivalence class of tuples $(F_0, \dots, F_m)$ as above, where two tuples are equivalent if they determine the same map.

Definition. A point $P \in V$ is a *regular point* of a rational map $\varphi \colon V \dashrightarrow \mathbb{P}^n$ if there is a representative $(F_0, \dots, F_m)$ of φ such that $F_j(P) \neq 0$ for some j. The *domain* of φ is the set of regular points of φ. A rational map φ is called a *morphism* if the domain of φ is V; in this case, we write $V \to \mathbb{P}^m$.

Example. A linear map $\varphi \colon \mathbb{P}^n \dashrightarrow \mathbb{P}^m$ is given by an $(m+1) \times (n+1)$ matrix $A = (a_{ij})$. Concretely, we can define $\varphi = (F_0, \dots, F_m)$ where $F_j = \sum_i a_{ij}X_i$. If A has rank $n+1 \leq m+1$, then φ is a morphism.

Example (projection from a point)**.** Let $P = (0 : 0 : 1) \in \mathbb{P}^2$. The *projection from P* is the rational map $\pi : \mathbb{P}^2 \dashrightarrow \mathbb{P}^1$ defined by $(a_0 : a_1 : a_2) \mapsto (a_0 : a_1)$. π is not regular at P, and regular everywhere else.

Let $C = \mathbb{V}(f_d)$ where f_d is a homogeneous polynomial of degree d. Suppose that $P \notin C$. The composition is therefore a morphism $\varpi : C \to \mathbb{P}^1$. One can show that for almost all choices of $Q \in \mathbb{P}^1$, the fibre $\varpi^{-1}(Q)$ is a set of size d.

Example. Let $C = \mathbb{V}(X_0X_2 - X_1^2) \subseteq \mathbb{P}^2$. Consider the projection from $(0 : 0 : 1)$, and restrict this projection to C to obtain a map $\pi : C \dashrightarrow \mathbb{P}^1$ defined by $\pi(a_0 : a_1 : a_2) = (a_0 : a_1)$. By changing representatives, we can show π is a morphism, even though $(0 : 0 : 1) \in C$.

The map π is determined by (X_0, X_1); we must look for other pairs (F_0, F_1) that determine the same rational map as π, so $F_0X_1 - F_1X_0 \in I^h(C) = (X_0X_2 - X_1^2)$. Notice that this relation is satisfied by (X_1, X_2), so π agrees with the function $\pi'(a_0 : a_1 : a_2) = (a_1 : a_2)$ on C. So π is regular at $(0 : 0 : 1) \in C$, so π is a morphism.

Observe that for $\pi : C \to \mathbb{P}^1$, $\pi^{-1}(q)$ is a single point for $q \in \mathbb{P}^1$. One can show also that π is surjective.

If W is a projective variety, a rational map (or morphism) $V \to W$ is a rational map (or morphism) $V \to \mathbb{P}^m$ with image contained in W. A morphism $\varphi : V \to W$ is an isomorphism if it has a two-sided inverse morphism.

Proposition. Let C be the vanishing locus of a homogeneous polynomial $f \in \mathbb{C}[X_0, X_1, X_2]$ of degree 2 in $\mathbb{P}^2$. Then, if f is irreducible then $C \simeq \mathbb{P}^1$.

Proof. By changing coordinates, we can assume $f = X_0X_2 - X_1^2$; the rank of the quadratic form is 2 as f is irreducible. By the example above, we have a morphism $\pi : C \to \mathbb{P}^1$ by projection from $(0 : 0 : 1)$. We define an inverse map $\mu : \mathbb{P}^1 \to \mathbb{P}^2$ by $\mu(Y_0 : Y_1) = (Y_0^2 : Y_0Y_1 : Y_1^2)$. The image of μ lies in C, and the compositions are inverse. □

There is only one conic in two-dimensional projective space, up to changing coordinates.

Example (Cremona transformation)**.** Consider the rational map $\mathbb{P}^2 \dashrightarrow \mathbb{P}^2$ given by

$$\kappa(X_0 : X_1 : X_2) = (X_1X_2 : X_0X_2 : X_0X_1)$$

This can be thought of as a coordinatewise reciprocal map. The Cremona transformation maps lines into conics. Suppose ℓ is a line not given by the vanishing locus of any of the coordinate functions X_i. Then, consider the subset $\kappa(\operatorname{dom}\kappa \cap \ell) \subseteq \mathbb{P}^2$; this is the analogue of the image in the case of rational maps. One can show that the closure of this set is a conic.

Example (Veronese embedding)**.** Let $F_0, \dots, F_m$ be the list of monomials of degree d in $X_0, \dots, X_n$, so $m = \binom{n+d}{d} - 1$. We define the $\nu_d : \mathbb{P}^n \to \mathbb{P}^m$ mapping $(\mathbf{a})$ to $(F_0(\mathbf{a}), \dots, F_m(\mathbf{a}))$. One can show this is a morphism. Note that the map $\mu(Y_0 : Y_1) = (Y_0^2 : Y_0Y_1 : Y_1^2)$ used in the previous proposition is an instance of this embedding. In general, ν_d is injective, and the image of ν_d is a projective variety isomorphic to $\mathbb{P}^n$. This fact has a straightforward but tedious proof.

Note that $\mathbb{P}^n \times \mathbb{P}^m \not\cong \mathbb{P}^{n+m}$.

Example (Segre embedding). Let $n, m > 0$ be integers. The *Segre embedding* is the map $\sigma_{mn} : \mathbb{P}^m \times \mathbb{P}^n \to \mathbb{P}^{mn+m+n}$ defined by $\sigma_{mn}((x_i), (y_j)) = (x_i y_j)$. We label the coordinates of $\mathbb{P}^{mn+m+n}$ using Z_{ij} for $0 \leq i \leq m$ and $0 \leq j \leq n$. Note that $(m+1)(n+1) - 1$; we have a map $U \times V \to U \otimes V$ and then take the projectivisation, giving the correct dimension.

Theorem. The map σ_{mn} is a bijection between $\mathbb{P}^m \times \mathbb{P}^n$ and the projective variety $\mathbb{V}(I)$ where I is the ideal generated by the $Z_{ij}Z_{pq} - Z_{iq}Z_{pj}$.

Proof. Clearly, $\sigma_{mn}(\mathbb{P}^m \times \mathbb{P}^n) \subseteq V = \mathbb{V}(I)$. Now, consider the affine piece $V_{00} = V \cap \{Z_{00} \neq 0\} \subseteq \mathbb{A}^{mn+m+n}$. The inhomogeneous ideal defining V_{00} is generated by $Y_{ij} - Y_{i0}Y_{0j}$ where $1 \leq i \leq m$ and $1 \leq j \leq n$, and $Y_{ij} = \frac{Z_{ij}}{Z_{00}}$. Note that elements $Y_{ij}Y_{pq} - Y_{iq}Y_{pj}$ for other indices automatically lie in this ideal. On this patch, σ_{mn} defines a morphism $\mathbb{A}^m \times \mathbb{A}^n \to \mathbb{V}(I_{00})$. There is an inverse $\mathbb{A}^{mn+m+n} \to \mathbb{A}^m \times \mathbb{A}^n$, given by

$$(Y_{ij}) \mapsto ((Y_{10}, \dots, Y_{m0}), (Y_{01}, \dots, Y_{0n}))$$

One can check that this is indeed an inverse; this process can be repeated for all other patches $\{Z_{ij} \neq 0\}$, so σ_{mn} is as claimed. □

Hence, if V, W are projective varieties, $V \times W$ is naturally also a projective variety.

3.6. Composition of rational maps

Let $\varphi : V \dashrightarrow W$ and $\psi : W \dashrightarrow Z$ be rational maps between irreducible varieties. The composition $\psi \circ \varphi$ of rational maps may not be well-defined, as the image of the domain of φ could lie entirely inside the locus of indeterminacy of ψ.

Definition. A rational map $\varphi : V \dashrightarrow W$ is *dominant* if $\varphi(\operatorname{dom}\varphi)$ is Zariski dense in W.

Proposition. If φ is dominant, then $\psi \circ \varphi$ is well-defined for any rational map $\psi : W \dashrightarrow Z$.

Proof. Let U denote a dense open set in $\operatorname{dom}\varphi$, and let U' be a dense open set in $\operatorname{dom}\psi$. Then, let $U'' = U \cap \varphi^{-1}(U')$, which is open in V. The composition $\psi \circ \varphi$ is well-defined on U''. This is a rational map as the composition of polynomials is a polynomial. □

Definition. If $\varphi : V \dashrightarrow W$ and $\psi : W \dashrightarrow V$ are such that $\varphi \circ \psi$ and $\psi \circ \varphi$ are equivalent to the identity map on W and V respectively, we say that V and W are *birational* and that φ and ψ are *birational maps*.

Example. Any isomorphism is birational.

Example. Consider the Cremona map $\kappa : \mathbb{P}^2 \dashrightarrow \mathbb{P}^2$ defined as above by $(x_0 : x_1 : x_2) \mapsto (x_1x_2 : x_0x_2 : x_0x_1)$. Intuitively, $(x_0 : x_1 : x_2) \mapsto \left(\frac{1}{x_0} : \frac{1}{x_1} : \frac{1}{x_2}\right)$. Then κ is self-inverse as a rational map, hence birational. It is not an isomorphism as it is not defined everywhere.

Remark. One can construct the group $\mathrm{Bir}(\mathbb{P}^2)$ of birational automorphisms of $\mathbb{P}^2$. This group contains a copy of $PGL_2(\mathbb{C})$ and the subgroup generated by $\varkappa$ above.

Theorem. Let V, W be irreducible projective varieties. Then V is birational to W if and only if $\mathbb{C}(V)$ and $\mathbb{C}(W)$ are isomorphic as fields.

Recall the similar result that if V, W are affine varieties, V is isomorphic to W if and only if $\mathbb{C}[V]$ and $\mathbb{C}[W]$ are isomorphic as $\mathbb{C}$-algebras.

Proof. Suppose first that V is birational to W, so $\varphi : V \dashrightarrow W$ is a birational map. Let $f \in \mathbb{C}(W)$. Then, f gives a function $W \dashrightarrow \mathbb{A}^1 = \mathbb{C}$, and composition gives a map of fields $\varphi^\star : \mathbb{C}(W) \to \mathbb{C}(V)$ defined by $f \mapsto f \circ \varphi$. Similarly, φ^{-1} gives a map $\mathbb{C}(V) \to \mathbb{C}(W)$, and the compositions are identical, so we obtain an isomorphism of fields.

For the converse, suppose we have $\mathbb{C}(V) \simeq \mathbb{C}(W)$ as fields. Suppose that $V \subseteq \mathbb{P}^n$ is not contained in $\{X_0 = 0\}$, and $W \subseteq \mathbb{P}^m$ is not contained in $\{Y_0 = 0\}$. We have shown that $\mathbb{C}(V) = \mathbb{C}(x_1, \dots, x_n)$ where x_i is the rational function determined by $\frac{X_i}{X_0}$. Similarly, $\mathbb{C}(W) = \mathbb{C}(y_1, \dots, y_m)$ where y_j is determined by $\frac{Y_j}{Y_0}$.

An isomorphism $\mathbb{C}(V) \simeq \mathbb{C}(W)$ identifies each y_j with $f_j(\mathbf{x})$ where f_j is a rational function in n variables. Writing each $f_j(\mathbf{x})$ as a rational function in the $\frac{X_i}{X_0}$, we can clear denominators by multiplying by some polynomial in the $\frac{X_i}{X_0}$ and homogenise with respect to X_0. We then obtain homogeneous polynomials $F_0, \dots, F_m$ in $X_0, \dots, X_n$ such that

$$f_j\left(\frac{X_1}{X_0}, \dots, \frac{X_n}{X_0}\right) = \frac{F_j(X_0, \dots, X_n)}{F_0(X_0, \dots, X_n)}$$

Now, $F_0, \dots, F_m$ determine a rational map $V \dashrightarrow W$. This can be repeated with the x_i and y_j reversed to obtain a rational map $W \dashrightarrow V$. One can show that these are inverses. □

4. Dimension

4.1. Tangent spaces

Let $V \subseteq \mathbb{A}^n$ be an affine hypersurface, so $V = \mathbb{V}(f)$. We assume that f is irreducible, so V is also irreducible. Let $P = (a_1, \dots, a_n) = (\mathbf{a}) \in V$. An affine line through P has the form $L = \{(a_1 + b_1 t, \dots, a_n + b_n t) \mid t \in \mathbb{C}\}$ for $(\mathbf{b}) \in \mathbb{C}^n \setminus \{\mathbf{0}\}$ is fixed.

The intersection $V \cap L$ is the set of points on L where f vanishes. This gives $0 = f(a_1 + b_1 t, \dots, a_n + b_n t) = g(t) = \sum_r c_r t^r$, a polynomial in t. Since $P \in V \cap L$, $c_0 = 0$. The linear term c_1 is given by $c_1 = \sum_i b_i \frac{\partial f}{\partial X_i}$. Geometric tangency of L to V is equivalent to the statement that $c_1 = 0$.

Definition. The line L through P is *tangent* to $V = \mathbb{V}(f)$ at P if it is contained in the *tangent space* of V at P, defined by $T^{\text{aff}}_{V,P} = \mathbb{V}(g) \subseteq \mathbb{A}^n$ where

$$g = \sum_{i=1}^n \left(\frac{\partial f}{\partial X_i}(P) \right)(X_i - a_i)$$

Note that g is linear. $T^{\text{aff}}_{V,P}$ is n-dimensional if $g = 0$ and $(n-1)$-dimensional otherwise, taking the dimensions as an affine space.

Definition. If $\dim T^{\text{aff}}_{V,P} = n$, we say that P is a *singular point* of V. Otherwise, it is a *smooth point*.

Example (nodal cubic)**.** Consider the affine hypersurface $C = \mathbb{V}(Y^2 - X^2(X+1))$. One can show by direct calculation that the only singular point is $(0,0)$.

Example (cusp)**.** Consider $C = \mathbb{V}(Y^2 - X^3)$. Again, the point $(0,0)$ is the only singular point.

Let $V \subseteq \mathbb{V}(F) \subseteq \mathbb{P}^n$ for F an irreducible homogeneous polynomial.

Definition. The *projective tangent space* to V at P is $T^{\text{proj}}_{V,P} = \mathbb{V}(G)$ where

$$G = \sum_{i=0}^n \left(\frac{\partial F}{\partial X_i}(P) \right) X_i$$

To see that $P \in \mathbb{V}(G)$, note that $F(X_0, \dots, X_n) = \frac{1}{\deg F} \sum_{i=0}^n X_i \frac{\partial F}{\partial X_i}$; this is sometimes known as *Euler's formula*. Smooth points and singular points are defined as in the affine case. From the inverse function theorem, if all points are smooth, the tangent space is a manifold.

The affine and projective tangent spaces are compatible in a particular sense. Let $V = \mathbb{V}(F) \not\subseteq \{X_0 = 0\}$, and consider $V_0 = V \cap U_0$. If $P \in V_0 \subseteq V$, we can compute $T^{\text{proj}}_{V,P} \cap U_0$ and $T^{\text{aff}}_{V_0,P}$, which are both subsets of $\mathbb{A}^n$. Let $V_0 = \mathbb{V}(f)$, then $F(X_0, \dots, X_n) = X_0^{\deg F} f\left(\frac{X_1}{X_0}, \dots, \frac{X_n}{X_0}\right)$. By computing $\frac{\partial F}{\partial X_i}$, we find that if $P \in V_0$, $T^{\text{proj}}_{V,P} \cap U_0 = T^{\text{aff}}_{V_0,P}$.

Proposition. The set of singular points on a nonempty irreducible projective hypersurface is a proper Zariski closed subset. In particular, the set of smooth points is dense.

Proof. The set of singular points is the intersection of V with $\bigcap_i \mathbb{V}\left(\frac{\partial F}{\partial X_i}\right)$, so is a closed subvariety of V. If $V \cap \bigcap_i \mathbb{V}\left(\frac{\partial F}{\partial X_i}\right) = V$, then by the Nullstellensatz, $\frac{\partial F}{\partial X_i} \in I^h(V)$. However, $I^h(V)$ is principal and generated by F. Since $\frac{\partial F}{\partial X_i}$ is homogeneous and of smaller degree, $\frac{\partial F}{\partial X_i} \mid F$ gives that $\frac{\partial F}{\partial X_i} = 0$. So F is a constant polynomial, giving $V = \mathbb{P}^n$ as it is nonempty, which has no singular points. □

We can extend the definition of a tangent space to general varieties not generated by a single polynomial.

Definition. Let $V \subseteq \mathbb{A}^n$ be an affine variety, and let $P \in V$. Then the *tangent space* to V at P is

$$T_{V,P} = \left\{ \mathbf{v} \in \mathbb{C}^n \,\middle|\, \sum_{i=1}^n v_i \frac{\partial f}{\partial x_i}(P) = 0 \text{ for all } f \in I(V) \right\} \subseteq \mathbb{C}^n$$

This is a vector subspace of $\mathbb{C}^n$.

Definition. Let $V \subseteq \mathbb{P}^n$ be a projective variety, and let $P \in V$. Suppose $V_j = V \cap \{X_j \neq 0\}$ is an affine piece containing P. Then the *tangent space* to V at P is $T_{V,P} = T_{V_j,P}$.

Note that this definition requires a choice of j; it is not clear that this is well-defined. This will be addressed by the following propositions.

Recall that $\mathbb{P}^n$ is covered by $U_0, \dots, U_n$, and $U_i \simeq \mathbb{A}^n$. Each point $P \in \mathbb{P}^n$ is contained in at least one of these U_i. If the index is unimportant, we will write $\mathbb{A}_n \subseteq \mathbb{P}^n$ rather than $U_i \subseteq \mathbb{P}^n$.

Let $V \subseteq \mathbb{P}^n, W \subseteq \mathbb{P}^m$ be irreducible varieties and $\varphi : V \dashrightarrow W$ be a rational map. Given $P \in \operatorname{dom} \varphi \subseteq V$ and $Q = \varphi(P) \subseteq W \cap \mathbb{A}^m$, we will now define $\mathrm{d}\varphi_P : T_{V,P} \dashrightarrow T_{W,P}$. Suppose φ is determined by $(F_0, \dots, F_m)$, where the F_i are homogeneous and of the same degree. By restricting to $\mathbb{A}^n$, we can write $\frac{F_j}{F_0}(1, X_1, \dots, X_n) = f_j \in \mathbb{C}(X_1, \dots, X_n)$. This gives rational functions $f_1, \dots, f_m$ on $V \cap \mathbb{A}^n$. The *derivative* of φ at P or *linearisation* of φ at P is defined by

$$\mathrm{d}\varphi_P(v) = \left(\sum_{i=1}^n v_i \frac{\partial f_j}{\partial X_i}(P) \right)_j$$

which is initially a function $T_{V,P} \to \mathbb{C}^m$. This can be thought of as an application of the matrix of derivatives of f at P to the vector v.

Proposition. (i) $\mathrm{d}\varphi_P(T_{V,P}) \subseteq T_{W,Q}$;

(ii) the linear map $\mathrm{d}\varphi_P$ depends only on φ and not the representatives;

(iii) if $\psi : W \dashrightarrow Z$ is rational with $\varphi(P) \in \operatorname{dom}\psi$, then $\mathrm{d}(\psi \circ \varphi)_P = \mathrm{d}\psi_{\varphi(P)} \circ \mathrm{d}\psi_P$;

(iv) if φ is birational and φ^{-1} is regular at $\varphi(P)$, then $\mathrm{d}\varphi_P$ is an isomorphism $T_{V,P} \simeq T_{W,Q}$.

Proof. Part (i). We use Y_j for coordinates in W. Replace V with V_0 and W with W_0. Let $g \in I(W)$, and consider its linearisation at Q. Applying the map $\varphi^\star$ on function fields, we obtain $\varphi^\star(g) = h = g(f_1, \dots, f_m) \in \mathbb{C}(V)$. Choose a representative in $\mathbb{C}(X)$, representing a rational function on V that is regular at P. This map vanishes when it is regular as $\varphi(\operatorname{dom}\varphi) \subseteq W$. By the chain rule,

$$\frac{\partial h}{\partial X_i}(P) = \sum_j \frac{\partial g}{\partial Y_j}(Q)\frac{\partial f_j}{\partial X_i}(P)$$

Hence, $v \in T_{V,P}$ gives $\mathrm{d}\varphi_P\,(v) \in T_{W,Q}$.

Part (ii). If $(F_0', \dots, F_m')$ is another representation of φ with corresponding rational functions $f_1', \dots, f_m' \in \mathbb{C}(V)$. Then $f_j - f_j'$ vanishes on V whenever it is defined, or equivalently, $f_j - f_j' = \frac{p_j}{q_j}$ where $p_j \in I(V)$ and $q_j(P) \neq 0$. Applying the quotient rule and the fact that $p_j \in I(V)$,

$$\frac{\partial(f_j - f_j')}{\partial X_i} = \frac{-1}{q_j(P)} = \frac{\partial p_j}{\partial X_i}(P) = 0$$

Hence, $v \in T_{V,P}$ gives $\sum_i v_i \frac{\partial(f_j - f_j')}{\partial X_i}(P) = 0$ as required.

Part (iii). Follows from the chain rule from multivariate calculus.

Part (iv). Immediate from (iii). □

Note that if $P \in U_i \cap U_j$, we have two different definitions of the tangent space $T_{V,P}$. Suppose that $V = \mathbb{P}^n$, then there is a birational map $p_{ij} : U_i \dashrightarrow U_j$ which is the identity on $U_i \cap U_j$. Part (iv) of the above proposition gives an isomorphism from T_{P,U_i} to T_{P,U_j} given by $\mathrm{d}p_{ij}$.

4.2. Smooth and singular points

Definition. Let V be an affine or projective variety. If V is irreducible, the *dimension* of V, written $\dim V$, is the minimum dimension of a tangent space for a point in V. If $P \in V$ and V is irreducible, we say P is a *smooth point* of V if $\dim T_{V,P} = \dim V$. Otherwise, P is a *singular point*. If V is reducible, we define $\dim V$ to be the maximum dimension of an irreducible component of V.

Theorem. Let V be a nonempty irreducible affine or projective variety. Then the set of smooth points of V is a nonempty open subset of V.

Proof. The fact that the set is nonempty is clear as the minimum dimension must be attained at a point. We can assume $V \subseteq \mathbb{A}^n$ is affine. If $P \in V$,

$$T_{V,P} = \left\{ \mathbf{v} \in \mathbb{C}^n \;\middle|\; \sum_{i=1}^n v_i \frac{\partial f_j}{\partial x_i}(P) = 0 \right\}$$

where f_j is some finite set of functions with $\mathbb{V}(\{f_j\}) = V$. Then

$$\dim T_{V,P} = n - \operatorname{rank} \frac{\partial f_j}{\partial X_i}(P)$$

For any $r \in \mathbb{N}$,

$$\{P \in V \mid \dim T_{V,P} \geq r\} = \left\{ P \in V \;\middle|\; \operatorname{rank} \frac{\partial f_j}{\partial X_i}(P) \leq n - r \right\}$$

This is the subvariety given by the vanishing locus of the $(n-r+1) \times (n-r+1)$ minors of this matrix $\frac{\partial f_j}{\partial X_i}(P)$, which is closed. □

Corollary. If V, W are irreducible and birational, then $\dim V = \dim W$.

4.3. Transcendental extensions

If $K \subseteq L$ are fields and $\alpha \in L$, we say that α is *transcendental* over K if it is not a solution to a nontrivial polynomial $f \in K[t]$. More generally, if $S \subseteq L$ is any set of elements, we say they are *algebraically independent* if they do not satisfy a nontrivial polynomial relation over K. A field extension $K/\mathbb{C}$ is a *pure transcendental extension* if K is generated by transcendental algebraically independent elements $x_1, \dots, x_n \in K$.

If V is an irreducible affine variety, recall that $\mathbb{C}(V) = FF\left(\mathbb{C}[\mathbf{X}]/_{I(V)}\right)$. If $V = \mathbb{P}^1$, $\mathbb{C}(V) \simeq \mathbb{C}(X)$.

Proposition. Let $K/\mathbb{C}$ be a finitely generated field extension. Then, there exists a pure transcendental subfield $K_0 = \mathbb{C}(x_1, \dots, x_m) \subseteq K$ such that K/K_0 is finite (and hence algebraic). Moreover, $K = K_0(y)$ for some $y \in K$.

Proof. The final statement follows from the primitive element theorem from Part II Galois Theory. We now prove the first part. K is finitely generated, so let $x_1, \dots x_n$ generate K. There is a maximal algebraically independent subset which after relabelling is given by $\{x_1, \dots, x_m\}$ for $m \leq n$. Then $x_{m+1}, \dots, x_n$ are algebraic over $K_0 = \mathbb{C}(x_1, \dots, x_m)$. □

Proposition. Let $K = \mathbb{C}(x_1, \dots, x_n)$, where $x_1, \dots, x_n$ are algebraically independent. Let x_{n+1} be algebraic over K. Define

$$I = \{g \in \mathbb{C}[X_1, \dots, X_{n+1}] \mid g(x_1, \dots, x_n, x_{n+1}) = 0\}$$

Then I is a principal ideal generated by an irreducible element $f \in \mathbb{C}[\mathbf{X}]$. Moreover, if f contains the variable X_i, then $\{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n, x_{n+1}\}$ is algebraically independent.

Proof. As $x_1, \dots, x_n$ are algebraically independent, the subring $R = \mathbb{C}[x_1, \dots, x_n] \subseteq K$ is isomorphic to the polynomial ring $\mathbb{C}[X_1, \dots, X_n]$. $\mathbb{C}[X_1, \dots, X_n]$ is a unique factorisation domain. There exist polynomials $g \in K[T]$ where x_{n+1} is a root, as it is algebraic. Since $K[T]$ is a principal ideal domain, the ideal of such polynomials is principal, and generated by a unique monic polynomial $h(t)$, called the minimal polynomial of x_{n+1}. The minimal polynomial is irreducible.

Let b be the least common multiple of the denominators in $h(t)$, so $b \in R$. By Gauss' lemma, $f = bh$ is irreducible in $R[T]$. By the isomorphism $R \simeq \mathbb{C}[X_1, \dots, X_n]$, we can think of f as an element of $\mathbb{C}[X_1, \dots, X_{n+1}]$.

We show that f generates I. Suppose $g \in \mathbb{C}[\mathbf{X}]$ such that $g(x_1, \dots, x_{n+1}) = 0$. In $K[T]$, $g(x_1, \dots, x_n, T)$ is divisible by $f(x_1, \dots, x_n)$. By Gauss' lemma, $f \mid g$ in $\mathbb{C}[\mathbf{X}]$. Hence f generates I as required. The last part is left as an exercise. □

Corollary. Let V be any irreducible variety. Then V is birational to a hypersurface.

Proof. Let K be the function field of V. By the above discussion, we can find elements that generate K that are given by $x_1, \dots, x_{n+1}$ where $x_1, \dots, x_n$ are algebraically independent and x_{n+1} is algebraic over $\mathbb{C}(x_1, \dots, x_n)$. By the previous proposition, $K \supseteq \mathbb{C}[x_1, \dots, x_{n+1}] = \mathbb{C}[X_1, \dots, X_{n+1}] / (f)$. We take the hypersurface $\mathbb{V}(f) \subseteq \mathbb{A}^{n+1}$. □

We have shown above that birational varieties have the same dimension. We therefore have the following corollary.

Corollary. Let W be an irreducible variety, and let $V = \mathbb{V}(f) \subseteq \mathbb{A}^n$ be an affine hypersurface birational to W, where f is non-constant. Then the dimension of W is equal to $n - 1$.

In the language of field theory, the dimension of W is the transcendence degree of the field $\mathbb{C}(W)$.

5. Algebraic curves

5.1. Curves

Definition. A *curve* is a variety of dimension 1.

For our purposes, a curve is taken to mean a smooth irreducible projective variety of dimension 1. By convention, a curve C implicitly has an expression as $\mathbb{V}(I) \subseteq \mathbb{P}^n$, but this ambient space will not play an important role.

Example. Let $f_d \in \mathbb{C}[X, Y, Z]$ be homogeneous of degree d. For almost all choices of coefficients, $\mathbb{V}(f_d)$ is a (smooth irreducible projective) curve. We will show that for $d, d' \geq 2$, $\mathbb{V}(f_d)$ and $\mathbb{V}(f_{d'})$ are never isomorphic.

Proposition. Let C be a curve, and let $D \subsetneq C$ be a proper Zariski closed subset. Then D is a finite union of points.

Proof. It suffices to prove this for irreducible affine curves $V \subseteq \mathbb{A}^n$. Let $W \subsetneq V$ be a proper irreducible closed subset; we will show this is a single point. By the Nullstellensatz, there is a strict containment $I(V) \subsetneq I(W)$.

If $t \in \mathbb{C}[W] \setminus \mathbb{C}$, we can use this to produce an element $y \in \mathbb{C}[V]$ as follows. $\varphi : W \hookrightarrow V$ gives the pullback map $\varphi^\star : \mathbb{C}[V] \to \mathbb{C}[W]$ which is a surjection. Take any y such that $\varphi^\star(y) = t$.

We can also take $x \in \mathbb{C}[V]$ such that $\varphi^\star(x) = 0$, so $x \notin \mathbb{C}$. One can show that x, y are algebraically independent in $\mathbb{C}(V)$, as t is transcendental. This gives two algebraically independent elements of $C(V)$, which has transcendence degree 1. So no such t can exist, so $\mathbb{C}[W] = \mathbb{C}$. Therefore W is a point. □

Recall that if V is an irreducible variety, it has a coordinate ring (if it is affine), a function field, a local ring at each point, and the maximal ideal of functions vanishing at the given point in the local ring. These can be specialised in the case of curves. Note that if C is a smooth irreducible projective curve, there exists $t \in \mathbb{C}(V)$ such that ${}^{\mathbb{C}(V)}\!/_{\mathbb{C}(t)}$ is finite.

Theorem. Let P be a smooth point of an irreducible curve V. Then, the ideal $\mathfrak{m}_P \trianglelefteq \mathcal{O}_{V,P}$ is principal.

A generator π_P of $\mathfrak{m}_P$ is called a *local parameter*, a *local coordinate*, or a *uniformiser*.

Proof. We assume P lies in the affine patch V_0 of V. By changing coordinates, we can set

$P = 0 \in \mathbb{A}^n$.

$$\begin{aligned}\mathbb{C}[V_0] &= \mathbb{C}[X_1, \dots, X_n]\big/_{I(V_0)} = \mathbb{C}[x_1, \dots, x_n];\\ \mathcal{O}_P &= \mathcal{O}_{V_0,P} = \left\{ \frac{f}{g} \,\middle|\, f, g \in \mathbb{C}[V_0], g \notin (x_1, \dots, x_n) \right\}\\ \mathfrak{m}_P &= \left\{ \frac{f}{g} \,\middle|\, f \in (x_1, \dots, x_n), g \notin (x_1, \dots, x_n) \right\} = x_1\mathcal{O}_P + \dots + x_n\mathcal{O}_P \subseteq \mathcal{O}_P\end{aligned}$$

where x_i is the image of X_i under the quotient map. More generally, if $J \trianglelefteq \mathcal{O}_P$ is any ideal, $\frac{f}{g} \in J$ if and only if $f \in J$. Therefore,

$$J = \left\{ \frac{f}{g} \,\middle|\, f \in J \cap \mathbb{C}[V_0], g \in \mathbb{C}[V_0], g(P) \neq 0 \right\}$$

In particular, J is finitely generated.

Since P is smooth, $T^{\mathrm{aff}}_{V_0,P}$ is a line, and by changing coordinates,

$$T_{V,P} = \{X_2 = X_3 = \dots = X_n = 0\}$$

We claim that x_1 generates $\mathfrak{m}_P$. Since $T_{V,P}$ is cut out by linearisations at $P = 0$ of elements in $I(V_0)$, there exist functions $f_2, \dots, f_n \in I(V_0)$ such that $f_j = X_j - h_j$ where h_j has no terms of degree less than 2. In $\mathcal{O}_P$,

$$x_j = h_j(x_1, \dots, x_n) \in (x_1^2, x_1x_2, \dots, x_n^2) = \mathfrak{m}_P^2$$

Thus, $\mathfrak{m}_P = \sum_{j=1}^n x_i\mathcal{O}_P = x_1\mathcal{O}_P + \mathfrak{m}_P^2$. The result that $\mathfrak{m}_P$ is generated by x_1 follows from Nakayama's lemma.

Lemma (Nakayama)**.** Let R be a ring, let M be a finitely generated R-module, and let $J \trianglelefteq R$ be an ideal. Then,

(i) if $JM = M$, there exists $r \in J$ such that $(1 + r)M = 0$; and

(ii) if $N \leq M$ is a submodule such that $JM + N = M$, then there exists $r \in J$ such that $(1 + r)M \subseteq N$.

Let

$$R = \mathcal{O}_L \supseteq J = \mathfrak{m}_P = M \supseteq N = (x_1)$$

and apply part (ii) of Nakayama's lemma to conclude. □

Corollary. Let $V = \mathbb{V}(f) \subseteq \mathbb{A}^2$ be an irreducible affine curve. Then, if $P \in V$ is a smooth point, the function $V \to \mathbb{C}$ defined by $Q \mapsto X(Q) - X(P)$ is a local parameter if and only if $\frac{\partial f}{\partial Y}(P) \neq 0$.

Proof. Use the proof of the above theorem. □

Corollary. Let P be a smooth point of a curve V. Then there exists a surjective group homomorphism $\nu_P : \mathbb{C}(V)^\star \to \mathbb{Z}$ called the *valuation* at P or *order of vanishing* at P, such that

(i) $\mathcal{O}_{V,P} = \{0\} \cup \{f \in \mathbb{C}(V)^\star \mid \nu_P(f) \geq 0\}$;

(ii) $\mathfrak{m}_p = \{0\} \cup \{f \in \mathbb{C}(V)^\star \mid \nu_P(f) > 0\}$;

(iii) if $f \in \mathbb{C}(V)^\star$, then for any local parameter π_P, we can write $f = \pi_P^{\nu_P(f)} u$ where $u \in \mathcal{O}_{V,P}^\star = \mathcal{O}_{V,P} \setminus \mathfrak{m}_P$.

We will 'filter' the ring $\mathcal{O}_{V,P}$ by ideals generated generated by powers π_P^k for $k \geq 0$.

Proof. We know that $\mathfrak{m}_P = (\pi_P)$, so $\mathfrak{m}_P^n = (\pi_P^n)$. Define $J = \bigcap_{n \geq 0} \mathfrak{m}_P^n$. Note that $J \trianglelefteq \mathcal{O}_{V,P}$ is a finitely generated ideal as we have seen in the previous proof, and moreover, $\mathfrak{m}_P J = \pi_P J = J$. By part (i) of Nakayama's lemma, it follows that $J = 0$. So only the zero function vanishes to infinite order.

For every $f \in \mathcal{O}_{V,P} \setminus \{0\}$, there exists a unique n such that $f \in \mathfrak{m}_P^n \setminus \mathfrak{m}_P^{n+1}$. Define $\nu_P(f) = n$ for this n. If $f \in \mathbb{C}(V) \setminus \mathcal{O}_{V,P} \setminus \{0\}$, we claim $f^{-1} \in \mathcal{O}_{V,P}$. Indeed, $f = \frac{g}{h}$ for $g, h \in \mathcal{O}_{V,P}$, so we can write $g = \pi_P^k u$ and $h = \pi_P^\ell u'$ where $k, l \geq 0$ and $u, u' \in \mathcal{O}_{V,P}^\star$. Since $f \notin \mathcal{O}_{V,P}$, it follows that $k < \ell$, so $f^{-1} \in \mathcal{O}_{V,P}$ as required. Given this, we can define $\nu_P(f) = -\nu_P(f^{-1})$ for such f.

As $\mathfrak{m}_P$ is a local ring, $\mathcal{O}_{V,P} \setminus \mathfrak{m}_P = \mathcal{O}_{V,P}^\star$, so every nonzero $f \in \mathbb{C}(V)$ is $\pi_P^{\nu_P(f)} u$ where $u \in \mathcal{O}_{V,P}^\star$, giving ν_P as desired. □

Example. Let $V = \mathbb{A}^1$ and $P = 0 \in \mathbb{A}^1$. Then

$$\mathcal{O}_{\mathbb{A}^1,0} = \left\{ \frac{f(t)}{g(t)} \,\middle|\, g(0) \neq 0 \right\}; \quad \mathfrak{m}_0 = \left\{ \frac{f(t)}{g(t)} \,\middle|\, f(0) = 0, g(0) \neq 0 \right\}$$

So $\mathfrak{m}_0$ is the set of $\frac{f(t)}{g(t)}$ where $t \mid f$. Then $\mathfrak{m}_0^k$ is the set of $\frac{f(t)}{g(t)}$ where $t^k \mid f$.

We can think of $\frac{f(t)}{g(t)}$ where $g(t) = a_0 + a_1 t + \dots + a_k t^k$ as $f(t)$ multiplied by the power series expansion of $g(t)^{-1}$ which has nonzero constant term. This product can be written as t^M multiplied by another power series with nonzero constant term. The valuation of f is $\nu_0\left(\frac{f}{g}\right) = M$.

Corollary. Let V be an irreducible curve and $f \in \mathbb{C}(V)$. If P is a smooth point, f or f^{-1} is regular at P.

Proof. f is regular at P if and only if $f \in \mathcal{O}_{V,P}$. The statement then follows by checking the sign of $\nu_P(f)$. □

Corollary. Let V be a smooth curve. Then any rational map $V \dashrightarrow \mathbb{P}^m$ is a morphism.

Proof. Reordering coordinates, we can assume the image of $\varphi : V \dashrightarrow \mathbb{P}^m$ is not contained in $\{X_0 = 0\}$. We write $\varphi = (G_0, \dots, G_m) = (1 : g_1 : \dots : g_m)$ where $g_j = \frac{G_J}{G_0} \in \mathbb{C}(V)$. If all $g_j \in \mathcal{O}_{V,P}$, the result holds. Otherwise, let $t = \min_j \{\nu_P(g_j)\}$, so $t < 0$. Note that $\min_j \{\nu_P(\pi_P^{-t} g_j)\} = 0$. Then $\varphi \sim (\pi_P^{-t} : \pi_P^{-t} g_1 : \dots : \pi_P^{-t} g_m)$ which is regular at P. □

Since every projective variety is contained in $\mathbb{P}^m$, any rational map from a curve to a projective variety is a morphism.

5.2. Maps between curves

Example. Let $C_d \subseteq \mathbb{P}^2$ be a smooth plane curve of degree d, so $C_d = \mathbb{V}(f)$ where f is homogeneous of degree d. Let $P \in \mathbb{P}^2$. Then, the projection from P, which is a rational map $\mathbb{P}^2 \dashrightarrow \mathbb{P}^1$, automatically restricts to a morphism $C_d \to \mathbb{P}^1$. This morphism is surjective, and most points in $\mathbb{P}^1$ have a fibre of size d.

Proposition. Let $\varphi : V \to W$ be a non-constant morphism of irreducible (possibly singular) projective curves. Then, for all $Q \in W$, the fibre $\varphi^{-1}(Q)$ is finite. The map φ induces an inclusion $\varphi^\star : \mathbb{C}(W) \hookrightarrow \mathbb{C}(V)$ which makes $\mathbb{C}(V)$ a finite extension of $\mathbb{C}(W)$.

Proof. For the first statement, $\varphi^{-1}(Q)$ is Zariski closed in V, so is either V or a finite set of points. As φ is not constant, the fibre is a finite set of points. V is infinite, so by the first part, $\varphi(V)$ is infinite and therefore dense in W. Since φ is dominant, $\varphi^\star$ is defined. The map is automatically injective. Let $t \in \mathbb{C}(W) \setminus \mathbb{C}$ with $\varphi^\star(t) = x$. Since $\mathbb{C}(V)$ has transcendence degree 1 over $\mathbb{C}$, $\mathbb{C}(V)$ is finite over $\mathbb{C}(x)$, so also over $\mathbb{C}(W)$. □

Definition. Let $\varphi : V \to W$ be a non-constant morphism of curves. The *degree* of φ is the degree of the field extension $\mathbb{C}(V)/\varphi^\star\mathbb{C}(W)$.

Definition. Let $\varphi : V \to W$ be a non-constant morphism of curves, let $P \in V$ be a smooth point, and define $Q = \varphi(P)$. We define the *ramification degree* of φ at P by $e_P = e(\varphi, P) = \nu_P(\varphi^\star \pi_Q)$, where π_Q is a local coordinate at Q.

Example. Consider the morphism $\varphi : \mathbb{A}^1 \to \mathbb{A}^1$ defined by $z \mapsto z^d$ for some $d \geq 1$. On rings, this is given by $\varphi^\star : \mathbb{C}[Y] \to \mathbb{C}[X]$ with $\varphi^\star(Y) = X^d$. On function fields, this map satisfies $\varphi^\star\mathbb{C}(Y) = \mathbb{C}(X^d)$, a subfield of $\mathbb{C}(X)$. The degree of φ is d. Let $P = 0 \in \mathbb{A}^1$, so $Q = 0 \in \mathbb{A}^1$. A local parameter near Q is Y, and $\varphi^\star(Y) = X^d$. $\nu_0(X^d) = d$, so the ramification degree of φ at 0 is d.

Now suppose $P = 1$, $\varphi(P) = Q = 1$. The local coordinate at Q is $Y - 1$. We can find $\nu_P(\varphi^\star(Y-1)) = 1$, so the ramification degree of φ at 1 is 1. Note that $\varphi^{-1}(1)$ is the set of dth roots of unity, which is a set of d points $R_1, \dots, R_d$. $\nu_{R_i}(\varphi^\star(Y-1)) = 1$ for each i.

Theorem. Let $\varphi : V \to W$ be a non-constant morphism of irreducible projective curves.

(i) φ is surjective.

(ii) Suppose V, W are smooth. Then, for any $Q \in W$, $\deg \varphi = \sum_{P \in \varphi^{-1}(Q)} e_P$.

(iii) At all but finitely many points $P \in V$, $e_P = 1$.

Definition. A *quasi-projective variety* U is a Zariski-open subset of a projective variety $V \subseteq \mathbb{P}^n$.

Example. All projective varieties are quasi-projective. All affine varieties are also quasi-projective. Products of affine and projective varieties are quasi-projective, such as $\mathbb{P}^n \times \mathbb{A}^m$. Note that rational functions, rational maps, morphisms, irreducibility, function fields, local rings, and other algebraic geometric concepts are defined for quasi-projective varieties in the same way.

Proposition (fundamental theorem of elimination theory)**.** The projection map $\mathbb{P}^n \times \mathbb{A}^m \to \mathbb{A}^m$ is a Zariski closed map.

Preimages and images of closed sets are closed under this map.

Remark. Consider the map $\pi : \mathbb{A}^2 \to \mathbb{A}^1$ given by projection onto the x-axis. Observe that π is not a closed map, as $\mathbb{V}(XY - 1)$ has image $\mathbb{A}^1 \setminus \{0\}$, which is not closed.

Given this proposition, we prove the following result.

Proposition. Let $\varphi : V \to W$ be a morphism of quasi-projective varieties. Suppose that V is projective. Then φ is closed.

Proof. Factorise φ as $V \to \Gamma_\varphi \subseteq V \times W \to W$, where $\Gamma_\varphi = \{(x, \varphi(x)) \mid x \in V\}$ is the graph of φ. Note that Γ_φ is closed as it is the preimage of the diagonal $\varphi \times \mathrm{id} : V \times W \to W \times W$. The diagonal $W \subseteq W \times W$ is closed, even though $W \times W$ is not given the product topology. Now, $V \subseteq \mathbb{P}^n$ is a closed subset as it is a projective variety. Hence, it suffices to show that the projection map $\mathbb{P}^n \times W \to W$ is closed. Moreover, if W is covered by affine varieties $\{U_i\}$, it further suffices to show that $\mathbb{P}^n \times U_i \to U_i$ is closed for all i. Any quasi-projective variety is covered by affine varieties as required. Finally, each U_i is a closed subset of $\mathbb{A}^m$ for some m with its subspace topology. It therefore suffices to show $\mathbb{P}^n \times \mathbb{A}^m \to \mathbb{A}^m$ is closed, which is the fundamental theorem of elimination theory. □

We can now prove part (i) of the above theorem. Part (ii) is nonexaminable, and part (iii) will be shown later.

Corollary. Let $\varphi : V \to W$ be a non-constant map between irreducible projective curves. Then φ is surjective.

Proof. The image of φ is closed, so either a finite set of points or W itself. Since it is non-constant, φ is surjective. □

Corollary. Let V be a smooth projective irreducible curve, and let $f \in \mathbb{C}(V)^\star$. Then,

(i) if f is regular at all points $P \in V$, then $f \in \mathbb{C}^\star$ is a constant;

(ii) the set of $P \in V$ such that $\nu_P(f) \neq 0$ is finite, and $\sum_{P \in V} \nu_P(f) = 0$.

Proof. Part (i). Given f, consider the morphism $\varphi = (1 : f) \colon V \to \mathbb{P}^1$. φ is a morphism because C is smooth. We want to find zeroes and poles of f. $\varphi(P) = (1 : 0)$ if and only if $f(P) = 0$, and $\varphi(P) = (0 : 1)$ if and only if f is not regular at P. This means that if f is everywhere regular, φ is not surjective, so it is constant.

Part (ii). We can assume f is non-constant. Let t denote the rational function $\frac{X_1}{X_0}$ on $\mathbb{P}^1$. By the pullback, we obtain $\varphi^\star t \in \mathbb{C}(V)$ is exactly $\frac{f}{1} = f$. For convenience, $(1 : 0) \in \mathbb{P}^1$ will be denoted 0, and $(0 : 1) \in \mathbb{P}^1$ will be denoted ∞.

Observe that t is a local parameter at $0 \in \mathbb{P}^1$, so if $f(P) = 0$, $e_P = \nu_P(\varphi^\star t) = \nu_P(f)$. Similarly, $\frac{1}{t} = \frac{X_0}{X_1}$ is a local parameter at $\infty \in \mathbb{P}^1$, so if $f(P) = \infty$, we have $e_P = \nu_P(\varphi^\star \frac{1}{t}) = -\nu_P(f)$. Finally, if $f(P) \neq 0, \infty$, then $\nu_P(f) = 0$. By the previous theorem, $\deg \varphi = \sum_{\varphi(P)=0} \nu_P(f) = \sum_{\varphi(P)=\infty} -\nu_P(f)$, giving the desired result. □

Hence, there are no non-constant holomorphic functions.

5.3. Divisors

We will only consider smooth projective irreducible curves from now on. Let V be a curve. There is a natural inclusion from the space of functions defined everywhere on V (isomorphic to $\mathbb{C}$) to the field of rational functions on V. However, this field $\mathbb{C}(V)$ is very large and difficult to study directly. The goal of divisor theory is to organise $\mathbb{C}(V)$ into manageable (finite-dimensional) pieces.

Note that a rational function $f \in \mathbb{C}(V)$ determines an open subset $U \subseteq V$ on which f is well-defined as a function $U \to \mathbb{C}$. For instance, we could define $U = V \setminus \{x \mid \nu_P(f) < 0\}$, which is V with a finite set of points removed. One idea is to study functions $f \in \mathbb{C}(V)$ that are well-defined away from a fixed set $\{P_1, \dots, P_n\}$.

Definition. A *divisor D* on a curve V is a finite formal linear combination $\sum_{P \in V} n_P[P]$, or equivalently, an element of the free abelian group $\bigoplus_{P \in V} \mathbb{Z}[P]$. If $D = \sum_{P \in V} n_P[P]$, its *degree* is $\deg D = \sum_{P \in V} n_P \in \mathbb{Z}$.

Note that $\deg \colon \mathrm{Div}(V) \to \mathbb{Z}$ is a group homomorphism. The kernel of deg is denoted $\mathrm{Div}^0(V)$. If $D = \sum n_P[P]$, we write $\nu_P(D) = n_P$.

Definition. Let $D \in \mathrm{Div}(V)$. The space of rational functions on V with poles bounded by D is

$$L(D) = \{f \in \mathbb{C}(V) \mid f = 0 \text{ or } \forall P \in V,\ \nu_P(f) + \nu_P(D) \geq 0\}$$

For instance, if $\nu_P(D) > 0$, f is allowed to have a pole at P of order at most $\nu_P(D)$. If $\nu_P(D) < 0$, f is forced to have a zero at P of order at least $-\nu_P(D)$.

Definition. Let $f \in \mathbb{C}(V)^\star$. The *divisor of f* is $\operatorname{div}(f) = \sum_{P \in V} \nu_P(f)[P]$.

Divisors of rational functions must have degree 0. Divisors of the form $\operatorname{div}(f)$ are called *principal divisors*. The set $\operatorname{Prin}(V)$ is the set of divisors $D \in \operatorname{Div}(V)$ such that $D = \operatorname{div}(f)$ for some $f \in \mathbb{C}(V)^\star$, and this is a subgroup of $\operatorname{Div}^0(V)$, as $\operatorname{div}(f \cdot g) = \operatorname{div} f + \operatorname{div} g$.

The quotient ${}^{\operatorname{Div}(V)}\!/\!_{\operatorname{Prin}(V)}$ is noted $\operatorname{Pic}(V) = \operatorname{Cl}(V)$, and this is called the *Picard group* or *class group* of V. The Picard group and class group coincide for smooth varieties, but are different in the study of general varieties and schemes.

Divisors D, D' are called *linearly equivalent* if $D - D'$ is $\operatorname{div}(f)$ for some $f \in \mathbb{C}(V)^\star$, so D is equivalent to D' in $\operatorname{Pic}(V)$. We write $D \sim D'$.

Proposition. Every degree 0 divisor on $\mathbb{P}^1$ is principal.

Note that every principal divisor is degree 0 in general.

Proof. Identify $\mathbb{P}^1$ with $\mathbb{C} \cup \{\infty\}$, where $\mathbb{C} \hookrightarrow \{(1 : z) \mid z \in \mathbb{C}\}$. Then, $D = \sum_{a \in \mathbb{C}} n_a[a] + n_\infty[\infty]$. Note that $n_\infty = -\sum_{a \in \mathbb{C}} n_a$. Let $f = \prod_{a \in \mathbb{C}} (t-a)^{n_a}$. This has a zero of order n_a at a. Hence, $\operatorname{div} f = D$; clearly, $\nu_a(\operatorname{div} f) = n_a$ for $a \in \mathbb{C}$, and $\frac{1}{t-a}$ is a local coordinate at ∞ for all $a \in \mathbb{C}$ where $t = \frac{X_1}{X_0}$, then we can calculate explicitly $\nu_\infty(\operatorname{div} f) = n_\infty$. □

It is not always the case that every degree 0 divisor on a curve V is principal and $\operatorname{Pic}(V)$ is nontrivial; this gives rise to the notion of genus.

Definition. Let $V \subseteq \mathbb{P}^n$ be a curve. Consider the hyperplane $\mathbb{V}(L) \subseteq \mathbb{P}^n$ where L is a homogeneous linear polynomial. Assume $V \not\subseteq \mathbb{V}(L)$. The *hyperplane section* of V by $\mathbb{V}(L)$ is $\operatorname{div} L = \sum_{P \in V} n_P[P]$, where if $X_i(P) \neq 0$, $n_P = \nu_P\left(\frac{L}{X_i}\right)$.

This is well-defined as $\nu_P\left(\frac{L}{X_i}\right) = \nu_P\left(\frac{L}{X_j}\right)$ for $X_i(p) \neq 0, X_j(P) \neq 0$, as $\frac{X_i}{X_j} \in \mathcal{O}_{V,P}^\star$ so $\nu_P\left(\frac{X_i}{X_j}\right) = 0$. Note that all n_P are nonnegative in this case.

Proposition. Let $V \subseteq \mathbb{P}^n$ be as above, and let L, L' be linear homogeneous polynomials, neither vanishing on V. Then there is an equality

$$\operatorname{div} L - \operatorname{div} L' = \operatorname{div}\left(\frac{L}{L'}\right)$$

In particular, $\operatorname{div} L - \operatorname{div} L'$ is principal, and $\deg \operatorname{div} L = \deg \operatorname{div} L'$.

Definition. Let $V \subseteq \mathbb{P}^n$ be a curve. Then the *degree* of V is $\deg \operatorname{div} L$ where $V \not\subseteq \mathbb{V}(L)$.

Remark. A line in $\mathbb{P}^2$ is degree 1. A conic is degree 2.

We can generalise these notions.

(i) If $\varphi : V \to \mathbb{P}^n$ is any non-constant morphism, and L is a linear form, we can similarly define $\operatorname{div} L$ by using $\sum_{P \in V} n_P[P]$ where $n_P = \nu_P\left(\frac{\varphi^\star L}{X_i}\right)$ where $X_i(P) \neq 0$. This generalises the case where φ is an inclusion. As before, we assume $\mathbb{V}(L)$ does not contain $\operatorname{Im}\varphi$. Note that this map need not be injective.

(ii) If G is homogeneous of degree $m \geq 1$ and $\varphi : V \to \mathbb{P}^n$, one can similarly define $\operatorname{div} G = \sum_{P \in V} n_P[P]$ where $n_P = \nu_P\left(\frac{\varphi^\star G}{X_i^m}\right)$ for any i such that $X_i(P) \neq 0$.

Theorem (weak form of Bézout's theorem). Let $V, V' \subseteq \mathbb{P}^2$ be smooth projective irreducible curves of degrees m, n. Then if $V \neq V'$, the number of intersection points of V and V' is at most mn.

We have already shown that this is the case when V' is a line on an example sheet.

Proof. Suppose V, V' are cut out by $\mathbb{V}(F), \mathbb{V}(G)$ of degrees m, n. We claim that the degree of $\operatorname{div} G$ as a divisor on V is mn. We can replace G by any other homogeneous polynomial of degree m by the previous proposition as it gives a linearly equivalent divisor. Replace G with L^m for a homogeneous linear polynomial L. Now, $\mathbb{V}(L) \cap V$ has size at most $n = \deg V$, so $\deg \operatorname{div} \varphi^\star G = nm$ as required, since $\operatorname{div}(\varphi^\star G) = \sum_{P \in V \cap \mathbb{V}(G)} n_P[P]$ where $n_P > 0$ (note that if $n_P > 0$ then G vanishes at P). □

5.4. Function spaces from divisors

Definition. A divisor D is called *effective* if $D = \sum n_P[P]$ for $n_P \geq 0$.

Recall that

$$L(D) = \{f \in \mathbb{C}(V) \mid f = 0 \text{ or } \operatorname{div} f + D \geq 0 \text{ pointwise}\}$$

is equivalently the set of $f \in \mathbb{C}(V)$ such that $\operatorname{div} f + D$ is effective.

Proposition. The set $L(D)$ is a complex vector subspace of $\mathbb{C}(V)$.

Proof. $\nu_P(f+g) \geq \min\{\nu_P(f), \nu_P(g)\}$, hence sums of the form $f+g$ lie in $L(D)$ if $f, g \in L(D)$. Clearly $L(D)$ is closed under scalar multiplication. □

Definition. Denote $\ell(D) = \dim_{\mathbb{C}} L(D)$.

Example. Let ∞ denote the point $(0 : 1) \in \mathbb{P}^1$, and let $D = m[\infty]$ where $m \geq 0$. Writing $t = \frac{X_1}{X_0}$, $L(D)$ is spanned by $1, t, t^2, \dots, t^m$. Hence, $\ell(D) = m + 1$.

Proposition. Let D be a divisor on V. Then,

(i) If $\deg D < 0$, then $L(D) = 0$.

(ii) If $\deg D \geq 0$, then $\ell(D) \leq \deg D + 1$.

(iii) For any $P \in V$, $\ell(D) \leq \ell(D - P) + 1$.

In particular, $L(D)$ is always finite-dimensional.

Proof. Part (i). If $L(D) \neq 0$ then there exists $f \neq 0$ with $f \in L(D)$ such that $\operatorname{div} f + D \geq 0$. But taking degrees, $\deg \operatorname{div} f = 0$ hence $\deg D \geq 0$, a contradiction.

Part (iii). Let $n = \nu_P(D)$. Define $\mathrm{ev}_P : L(D) \to \mathbb{C}$ by $f \mapsto (\pi_P^n f)(P)$, intuitively extracting the first nonzero term of the power series defining f at P. The kernel of this homomorphism is $L(D - P)$.

Part (ii). This now follows from parts (i) and (iii). If $d = \deg D$, then $\ell(D) \leq \ell(D - (d + 1)P) + d + 1 = d + 1$ where the latter equality holds as $\deg(D - (d + 1)P) < 0$. □

Proposition. Let D, E be divisors on a curve V such that $D \sim E$, or equivalently, $D - E$ is principal. Then $L(D)$ and $L(E)$ are isomorphic as complex vector spaces. In particular, $\ell(D) = \ell(E)$.

Proof. If $D - E$ is principal, it can be written as $\operatorname{div}(g)$. Multiplication by g (respectively g^{-1}) gives a linear map (respectively its inverse) $L(D) \to L(E)$. □

6. Differentials

6.1. Differentials over fields

Differentials on curves will allow us to tackle some interesting questions.

(i) Given $D \in \operatorname{Div}(V)$, can we calculate (or bound) $\ell(D)$?

(ii) (Brill–Noether theory) For what integers r, d does a curve V admit a morphism $\varphi : V \to \mathbb{P}^r$ of degree d such that $\operatorname{Im} V$ is not contained in a hyperplane?

(iii) (Hurwitz problem) When does there exist a morphism $V \to W$ of smooth projective curves?

Definition. Let $K/\mathbb{C}$ be a field extension. The *space of differentials*, written $\Omega_{K/\mathbb{C}}$, is the quotient vector space ${}^{M}\!/_{N}$ where M is the K-vector space spanned by symbols δx where $x \in K$, and N is the subspace of M generated by

$$\delta(x+y) - \delta(x) - \delta(y); \quad \delta(xy) - x\delta(y) - y\delta(x); \quad \delta(a)$$

where $x, y \in K, a \in \mathbb{C}$. Given $x \in K$, we define $\mathrm{d}x = \delta x + N \in \Omega_{K/\mathbb{C}}$. The *exterior derivative* is the $\mathbb{C}$-linear map $\mathrm{d} : K \to \Omega_{K/\mathbb{C}}$ mapping x to $\mathrm{d}x$.

Remark. More generally, if $\varphi : A \to B$ is a ring homomorphism, we could have defined $\Omega_\varphi = \Omega_{B/A}$ as a B-module as above.

Definition. Let U be a K-vector space. A $\mathbb{C}$-linear transformation $D : K \to U$ is called a *derivation* if $D(xy) = xD(y) + yD(x)$.

Example. The map $\mathrm{d} : K \to \Omega_{K/\mathbb{C}}$ is a derivation. The map $\frac{\mathrm{d}}{\mathrm{d}x} : \mathbb{C}(X) \to \mathbb{C}(X)$ is a derivation.

Lemma (universal property). Let U be a K-vector space A map $D : K \to U$ is a derivation if and only if there is a K-linear map $\lambda : \Omega_{K/\mathbb{C}} \to U$ such that $\lambda(\mathrm{d}x) = D(x)$ for all $x \in K$.

$$\begin{tikzcd}
 & K \arrow[dl, "d"'] \arrow[dd, "D"] \\
\Omega_{K/\mathbb{C}} \arrow[dr, dashed, "\lambda"'] & \\
 & U
\end{tikzcd}$$

The proof is very simple and omitted. Intuitively, $\mathrm{d} : K \to \Omega_{K/\mathbb{C}}$ is the 'best possible' derivation.

Remark. For any derivation D, if $y \in K$ and $y \neq 0$, $D(x) = D\left(y \cdot \frac{x}{y}\right) = yD\left(\frac{x}{y}\right) + \frac{x}{y}D(y)$, giving the quotient rule.

$$D\left(\frac{x}{y}\right) = \frac{yDx - xDy}{y^2}$$

Lemma. (i) Let $f = \frac{h}{g} \in \mathbb{C}(X_1, \dots, X_n)$ and write $y = f(x_1, \dots, x_n)$ for $x_1, \dots, x_n \in K$. Then

$$\mathrm{d}y = \sum_{i=1}^{n} \frac{\partial f}{\partial X_i}(x_1, \dots, x_n)\,\mathrm{d}x_i$$

(ii) If $K = \mathbb{C}(x_1, \dots, x_n)$ for $x_i \in K$, then $\Omega_{K/\mathbb{C}}$ is spanned by $\mathrm{d}x_1, \dots, \mathrm{d}x_n$ as a K-vector space.

Proof. Part (i) follows from the rules of calculus for $\mathrm{d}(xy)$, $\mathrm{d}\left(\frac{x}{y}\right)$ and $\mathbb{C}$-linearity. Part (ii) is immediate from part (i). □

We have obtained divisors in two different ways: from rational functions, and from hyperplane sections of $V \to \mathbb{P}^r$. We will do the reverse, we will first construct divisors, and then use them to build maps $V \to \mathbb{P}^r$. Differentials are another way to construct divisors.

From now, we will write Ω_K for $\Omega_{K/\mathbb{C}}$.

Theorem. Let $K/\mathbb{C}(t)$ be finite, with t transcendental over $\mathbb{C}$. Then Ω_K is one-dimensional as a K-vector space, and is spanned by $\mathrm{d}t$.

Proof. First, suppose $K = \mathbb{C}(t)$, the function field of $\mathbb{P}^1$. By the lemma above, Ω_K is spanned by $\mathrm{d}t$. We need to show that Ω_K is nonzero, then it is clearly one-dimensional. By the universal property, it suffices to exhibit a single nonzero derivation on K. The function $\frac{\mathrm{d}}{\mathrm{d}t}$ is one such derivation.

Now suppose $K \neq \mathbb{C}(t)$. Write $K_0 = \mathbb{C}(t)$, so $K = \mathbb{C}(t, \alpha) = K_0(\alpha)$ for $\alpha \in K \setminus K_0$ algebraic over K_0. Let $h(t) \in K_0[X]$ be the minimal polynomial of α. By minimality of h, $h'(\alpha) \neq 0$ as it does not have a double root. By the previous lemma, $\mathrm{d}t, \mathrm{d}\alpha$ span Ω_K as a K-vector space.

If $f \in K_0[X]$, write $D_t f$ for $\frac{\partial f}{\partial t}$, by t-differentiating the coefficients. The lemma gives $0 = \mathrm{d}(h(\alpha)) = D_t h(\alpha)\,\mathrm{d}t + h'(\alpha)\,\mathrm{d}\alpha$. Hence Ω_K is spanned by $\mathrm{d}t$, so it suffices to show Ω_K is nonzero. As in the first part, it suffices to exhibit a single nonzero derivation on K.

First, define $D \colon K_0[X] \to K$ by $D(f) = D_t f$ if $f \in K_0$, $D(X) = \frac{-(D_t h)(\alpha)}{h'(\alpha)}$, and $D(X^n) = n\alpha^{n-1}D(X)$. One can check that the ideal $hK_0[X]$ is mapped to zero under D. This exhibits a nonzero derivation as required. □

6.2. Rational differentials

Definition. Denote $\Omega_V = \Omega_{\mathbb{C}(V)/\mathbb{C}}$. Elements of Ω_V are called *rational differentials.* A differential $\omega \in \Omega_V$ is *regular* at a point $P \in V$ if ω can be expressed as $\sum_i f_i\,\mathrm{d}g_i$ where $f_i, g_i \in \mathcal{O}_{V,P}$. Write

$$\Omega_{V,P} = \{\omega \in \Omega_V \mid \omega \text{ regular at } P\} \subseteq \Omega_V$$

Note that $\Omega_{V,P}$ is not a vector subspace over $\mathbb{C}(V)$, since we can multiply by functions that are not regular. However, it is a module over $\mathcal{O}_{V,P}$.

Recall that $\mathcal{O}_{V,P}$ contains the maximal ideal $\mathfrak{m}_P$, which is principal, giving local coordinates. We can make a similar construction in the context of differentials.

Theorem. $\Omega_{V,P}$ is a free $\mathcal{O}_{V,P}$-module generated by $\mathrm{d}\pi_P$ where π_P is a local coordinate at P. In other words, $\Omega_{V,P} = \left\{f \, \mathrm{d}\pi_P \mid f \in \mathcal{O}_{V,P}\right\}$.

Remark. If π, π' are local coordinates at P, $\mathrm{d}\pi = u \, \mathrm{d}\pi'$ where $u \in \mathcal{O}^\star_{V,P}$. More generally, if $\omega \in \Omega_V$, then $\pi^j \omega$ is regular, so lies in $\Omega_{V,P}$, for sufficiently large k. Given this theorem, we can always write $\omega \in \Omega_V$ as $f \, \mathrm{d}\pi_P$ where π_P is a local coordinate at P and $f \in \mathbb{C}(V)$.

Definition. Let $\omega \in \Omega_V$ and $P \in V$. Define $\nu_P(\omega) = \nu_P(f)$ where $\omega = f \, \mathrm{d}\pi_P$ and π_P is a local coordinate at P.

Lemma. Let $\omega \in \Omega_V$ be a nonzero differential. Then, $\nu_P(\omega) \neq 0$ for only finitely many points P.

Proof. As $\nu_P(f \, \mathrm{d}g) = \nu_P(f) + \nu_P(\mathrm{d}g)$ and $\nu_P(f) = 0$ for all but finitely many points, it suffices to only prove this lemma for the case $\omega = \mathrm{d}g$. Moreover, as g must be non-constant as $\mathrm{d}g \neq 0$, we can assume that g is transcendental. hence, ${}^{\mathbb{C}(V)}\!/\!_{\mathbb{C}(g)}$ is a finite extension. Consider $(1 : g) \colon V \to \mathbb{P}^1$. By the finiteness theorem for rational functions, there are only finitely many $P \in V$ such that $g(P) = \infty$ or $e_P > 1$.

If P is a point where $e_P = 1$, so the function is unramified, $\varphi^\star(t - g(P))$ is a local coordinate at P. But $\varphi^\star(t - g(P))$ is $g - g(P)$, so $\nu_P(\mathrm{d}g) = 0$. □

Differentials provide another source of divisors.

Definition. Let $\omega \in \Omega_V$. Then $\operatorname{div} \omega = \sum_{P \in V} \nu_P(\omega)[P]$.

Proposition. Let ω, ω' be nonzero rational differentials on V. Then, $\operatorname{div}\omega - \operatorname{div}\omega'$ is principal.

Proof. Since Ω_V is one-dimensional over $\mathbb{C}(V)$, we can write $\omega = f\omega'$ where $f \in \mathbb{C}(V)$. It follows from the definitions that $\operatorname{div}\omega - \operatorname{div}\omega' = \operatorname{div} f$. □

If ω is a nonzero differential, $\operatorname{div}\omega$ gives a well-defined element in $\operatorname{Pic}(V) = \operatorname{Cl}(V) = {}^{\operatorname{Div}(V)}\!/\!_{\operatorname{Prin}(V)}$. We say that $\operatorname{div}\omega$ is a *canonical divisor*, and its equivalence class is the *canonical class*, denoted K_V. Sometimes K_V is also simply called the canonical divisor.

We now prove the above theorem.

Proof. We want to check that $\mathrm{d}\pi_P$ generates the module $\Omega_{V,P}$ over $\mathcal{O}_{V,P}$. Clearly $\mathcal{O}_{V,P} \, \mathrm{d}\pi_P \subseteq \Omega_{V,P}$; we want to check that the converse holds. Given $f \in \mathcal{O}_{V,P}$, $f - f(P) \in \mathfrak{m}_P$. Hence,

$f = f(P) + \pi_P g \in \mathcal{O}_{V,P}$ where $g \in \mathcal{O}_{V,P}$. By the Leibniz rule, $\mathrm{d}f = g\,\mathrm{d}\pi_P + \pi_P\,\mathrm{d}g \in \mathcal{O}_{V,P}\,\mathrm{d}\pi_P + \pi_P\Omega_{V,P}$. Assume that $\Omega_{V,P}$ is finitely generated. Observe that

$$\mathcal{O}_P\,\mathrm{d}\pi_P \subseteq \Omega_{V,P} \subseteq \mathcal{O}_P\,\mathrm{d}\pi_P + \pi_P\Omega_{V,P}$$

Apply Nakayama's lemma to $R = \mathcal{O}_{V,P}, J = \mathfrak{m}_P, M = \Omega_{V,P}, N = \mathcal{O}_{V,P}\,\mathrm{d}\pi_P$.

To show $\Omega_{V,P}$ is finitely generated, choose an affine patch $V_0 \subseteq V$ containing P. Then $C[V_0] = \mathbb{C}[x_1, \dots, x_n]$ where the x_i generate $\mathbb{C}[V_0]$. If $f \in \mathcal{O}_{V,P}$, we can write $f = \frac{g}{h}$ where g, h are polynomials and $h(P) \neq 0$. Thus

$$\mathrm{d}f = \sum_{i=1}^{n} \left(\frac{h\frac{\partial g}{\partial X_i} - g\frac{\partial h}{\partial X_i}}{h^2} \right)(x_1, \dots, x_n)\,\mathrm{d}x_i$$

But $h(P) \neq 0$, so $\mathrm{d}f$ is in the $\mathcal{O}_{V,P}$-span of $\{\mathrm{d}x_i\}$. □

Example. Let $V = \mathbb{P}^1$, and let t be the coordinate on the standard $\mathbb{A}^1 \subseteq \mathbb{P}^1$. For any $a \in \mathbb{C}$, the rational function $(t-a)$ is a local coordinate. At infinity, $\frac{1}{t}$ is a local coordinate.

We now calculate $\operatorname{div}\mathrm{d}t$. We have $\nu_a(\mathrm{d}t) = \nu_a(\mathrm{d}(t-a)) = 0$ for all $a \in \mathbb{C}$. Note that $\mathrm{d}t = -t^2\,\mathrm{d}\left(\frac{1}{t}\right)$ so

$$\nu_\infty(\mathrm{d}t) = \nu_\infty\left(\frac{-1}{\left(\frac{1}{t}\right)^2}\,\mathrm{d}\left(\frac{1}{t}\right) \right) = -2$$

Hence $\operatorname{div}\mathrm{d}t = -2[\infty]$, so the degree is nonzero, hence this divisor is not principal.

Definition. Let V be a curve. The *genus* of V is $g(V) = \ell(K_V)$.

$L(K_V)$ is not well-defined, but $\ell(K_V)$ is. Note that if $V = \mathbb{P}^1$, then $\operatorname{div}\mathrm{d}t = -2[\infty]$, so $\ell(K_{\mathbb{P}^1}) = 0$, as there are no rational functions on $\mathbb{P}^1$ that vanish to order 2 at infinity, apart from the zero function.

6.3. Differentials on plane curves

We will study curves in $\mathbb{P}^2$.

Example (smooth plane cubics). Consider $V = \mathbb{V}(F) \subseteq \mathbb{P}^2$ where $F = X_0X_2^2 - \prod_{i=1}^{3}(X_1 - \lambda_i X_0)$ with $\lambda_1, \lambda_2, \lambda_3$ distinct complex numbers. This curve is nonsingular. To calculate the genus, we take the following steps.

(i) We first use the affine equation $f(x,y) = y^2 - \prod_{i=1}^{3}(x - \lambda_i)$, and write $f(x,y) = y^2 - g(x,y)$. Differentiating, $2y\,\mathrm{d}y = g'(x)\,\mathrm{d}x$ is a nontrivial relation in Ω_V.

(ii) Using this relation, we choose a convenient differential $\omega \in \Omega_V$; in this case, we will choose $\omega = \frac{\mathrm{d}x}{y}$.

(iii) Calculate $\operatorname{div}\omega$ by using the fact that if $\frac{\partial f}{\partial y}(P)$ is nonzero, $x - x(P)$ is a local parameter, and if $\frac{\partial f}{\partial x}(P)$ is nonzero, $y - y(P)$ is a local parameter.

We find that $K_V = 0$. Hence, $g(V) = 1$ as $\ell(0) = 1$.

Theorem. Let V be a smooth plane cubic. Then $g(V) = 1$, and in particular, $V \not\simeq \mathbb{P}^1$.

Proof. Change coordinates into the example above. □

Theorem. Let $V = \mathbb{V}(F) \subseteq \mathbb{P}^2$ be a smooth projective plane curve of degree d. Then $K_V = (d-3)H$ where H is the divisor class associated to a hyperplane section of V.

Proof. First, we will select a differential $\omega \in \Omega_V$. Change coordinates such that $(0 : 1 : 0) \notin V$. Let $x = \frac{X_1}{X_0}, y = \frac{X_2}{X_0}$ be elements of $\mathbb{C}(V)$. Set $f(X,Y) = F(1,X,Y)$, so $f(x,y) = 0$ in $\mathbb{C}(V)$. Differentiating, $\frac{\partial f}{\partial X}(x,y)\,\mathrm{d}x + \frac{\partial f}{\partial Y}(x,y)\,\mathrm{d}y = 0$ is a relation in Ω_V. Choose

$$\omega = \frac{\mathrm{d}x}{\frac{\partial f}{\partial Y}(x,y)} = \frac{-\,\mathrm{d}y}{\frac{\partial f}{\partial X}(x,y)}$$

Then, we will calculate $\operatorname{div}\mathrm{d}\omega$ in this affine patch. If $\frac{\partial f}{\partial Y}(P) \neq 0$, then $x - x(P)$ is a local coordinate at P. Then, $\nu_P(\omega) = \nu_P\left(\frac{1}{\frac{\partial f}{\partial Y}}(x,y)\right) = 0$. Otherwise, $\frac{\partial f}{\partial X}(P) \neq 0$ by smoothness, so $y - y(P)$ is a local coordinate and $\nu_P(\omega) = 0$.

Since $(0 : 1 : 0) \notin V$, any point at infinity in V is not contained in $\{X_2 = 0\}$. The equation for V on the patch $\{X_2 \neq 0\}$ is $g(z,w) = 0$ where $z = \frac{X_0}{X_2} = \frac{1}{y}$ and $y = \frac{X_1}{X_2} = \frac{x}{y}$ and $g(Z,W) = F(Z,W,1)$ in $\mathbb{C}[Z,W]$. Select a different differential

$$\eta = \frac{\mathrm{d}z}{\frac{\partial g}{\partial W}(z,w)} = \frac{-\,\mathrm{d}w}{\{g\}Z(z,w)}$$

By the same argument as before, $\nu_P(\eta) = 0$ for all P in the patch $\{X_2 \neq 0\}$. Using $f(X,Y) = Y^d g\left(\frac{1}{X}, \frac{X}{Y}\right)$ and differentiating, we find $\omega = Z^{d-3}\eta$. If $X_2(P) \neq 0$, we calculate $\nu_P(\omega) = (d-3)\nu_P(z) + \nu_P(\eta) = (d-3)\nu_P(z)$. As $Z = \frac{X_0}{X_2}$, $\operatorname{div}\omega = (d-3)\operatorname{div}X_0$ as claimed. □

Proposition. If $f(x,y) = 0$ is an affine patch equation for a smooth projective plane curve, and $\deg f \geq 3$, then

$$\left\{ \frac{x^r y^s\,\mathrm{d}x}{\frac{\partial f}{\partial y}} \,\middle|\, 0 \leq r,s;\ r+s \leq d-3 \right\}$$

is a basis for $L(K_V)$ for the representative of K_V given by $(d-3)H$ where H is the hyperplane at infinity.

The dx term can be considered a dummy symbol, meant to indicate that we think of the term as a differential.

Proof. The proof is non-examinable, and follows from the same argument as the proof of the previous theorem. □

Corollary. If $d, d' \geq 2$ are distinct integers, then smooth plane curves of degrees d, d' are never isomorphic.

Proof. $\deg K_V$ depends only on V up to isomorphism. □

In particular, there are infinitely many distinct curves up to isomorphism.

6.4. The Riemann–Roch theorem

Theorem. Let V be a smooth irreducible projective curve of genus g, and let D be a divisor on V. Let K_V be the canonical divisor class. Then,

$$\ell(D) - \ell(K_V - D) = \deg(D) - g + 1$$

The proof is beyond the scope of this course. This theorem is related to Stokes' theorem and the Gauss–Bonnet theorem.

Corollary. Let K be the canonical divisor on V. Then, $\deg(K) = 2g - 2$.

Note that $2g - 2 = -\chi(V)$, the negative of the Euler characteristic of V.

Proof. Let $D = K$ in the Riemann–Roch theorem, and use $\ell(0) = 1$. □

Corollary. Let V be a smooth projective plane curve of degree d. Then the genus is $g(V) = \frac{(d-1)(d-2)}{2}$.

Proof. We have seen that if $d = 1, 2$ then $V \simeq \mathbb{P}^1$. If $d \geq 3$, we have seen that K is linearly equivalent to $(d-3)H$ where H is a hyperplane section. But $\deg(H) = d$, hence the result follows from the Riemann–Roch theorem. □

Given a divisor D on V, calculating $\ell(D)$ is hard with the techniques discussed so far. However, the Riemann–Roch theorem can be used to compute this for most D.

Corollary. If $\deg(D) > 2g - 2$, then $\ell(D) = \deg(D) - g + 1$.

Proof. The divisor $K - D$ has negative degree, hence $\ell(K - D) = 0$. □

We can compare this to the case $V = \mathbb{P}^1$, where we saw by direct calculation that $\ell(D) = \deg(D) + 1$.

Corollary. Suppose $g(V) = 1$. Then if D is a divisor with $\deg(D) > 0$, then $\ell(D) = \deg(D)$.

Proof. $\ell(K - D) = \ell(-D) = 0$. □

Let V be a curve of genus 1, and fix $P_0 \in V$. Let $P, Q \in V$, then $P + Q - P_0$ is equivalent to a unique effective divisor of degree 1. So $P + Q - P_0$ is equivalent to R for a unique $R \in V$. Indeed, $\deg(P + Q - P_0) = 1$ hence $\ell(P + Q - P_0) = 1$, so there exists a function $f \in \mathbb{C}(V)$ such that $(P + Q - P_0) + \operatorname{div}(f)$ is effective, and hence equal to a point R. It is unique as $\ell(P + Q - P_0) = 1$, and scalar multiples of f give the same divisor.

In other words, given $E = (V, P_0)$ as above, we can define $P +_E Q = R$ using the preceding notation. The pair (V, P_0) where $g(V) = 1, P_0 \in V$ is called an *elliptic curve*. Topologically, such V in the Euclidean topology are homeomorphic to $\mathbb{S}^1 \times \mathbb{S}^1$; the group law defined by $+_E$ and that defined on $\mathbb{S}^1 \times \mathbb{S}^1$ in fact coincide.

Theorem. The operation $+_E$ gives E the structure of an abelian group with identity P_0. Moreover, the map $E \to \mathrm{Cl}^0(E) = \mathrm{Cl}^0(V)$ defined by $P \mapsto [P - P_0]$ is an isomorphism of groups.

Proof. Let $\beta(P) = [P - P_0] \in \mathrm{Cl}^0(E) = {}^{\mathrm{Div}^0(E)}\!/\!_{\mathrm{Prin}(E)}$. First, we show injectivity. Suppose $\beta(P) = \beta(Q)$, so $P - P_0 \sim Q - P_0$, where $\sim$ denotes linear equivalence. Hence $P \sim Q$. However, $\ell(P) = 1$ by the Riemann–Roch theorem, so $P = Q$.

Now, we show surjectivity. Suppose D has degree 0. We want to show D is equivalent to $[P - P_0]$ for some P. Since the degree of $D + P_0$ is 1, $\ell(D + P_0) = 1$ by Riemann–Roch. Hence there exists $P \in V$ such that $D + P_0 \sim P$. So $D = \beta(P)$ as required.

Hence β is a bijection of sets, so it remains to check that β is a homomorphism; this is straightforward. □

Theorem. Let $E = (V, P_0)$ be the elliptic curve given by $\mathbb{V}(F)$ where $F = X_0 X_2^2 - \prod_{i=1}^3 (X_1 - \lambda_i X_0)$. Choose $P_0 = (0 : 0 : 1)$. Then, $P +_E Q +_E R = 0_E$ if and only if P, Q, R are collinear in $\mathbb{P}^2$.

The proof is nonexaminable.

Given a morphism $\varphi : V \to W$ of curves, we wish to understand the relation between $g(V)$ and $g(W)$. Let $\omega = f\,\mathrm{d}t \in \Omega_W$, where $\mathbb{C}(W)/\mathbb{C}(t)$ is finite. Since $\mathbb{C}(V)/\mathbb{C}(t)$ is finite, Ω_V is generated by $\mathrm{d}\varphi^\star t$. Define the pullback $\Omega_W \to \Omega_V$ by $\varphi^\star\omega = \varphi^\star f\,\mathrm{d}\varphi^\star t$. Let P be a point on V, and $Q = \varphi(P)$. We compare $\nu_P(\varphi^\star\omega)$ and $\nu_Q(\omega)$.

Lemma. Let π_P, π_Q be local parameters at P, Q. Let e_P be the ramification degree at P, so $\varphi^\star(\pi_Q) = u\pi_P^{e_P}$ where u is a unit in $\mathcal{O}_{V,P}$. Then, $\nu_P(\varphi^\star(\mathrm{d}\pi_Q)) = e_P - 1$. More generally, $\nu_P(\varphi^\star\omega) = e_P \nu_Q(\omega) + e_P - 1$.

This can be thought of as a generalisation of the rule $\frac{\mathrm{d}}{\mathrm{d}x}\{x^n\} = nx^{n-1}$.

Proof. For the first part, we have that $\varphi^\star(\pi_Q) = u\pi_P^{e_P}$, so differentiating and taking valuation gives the desired result. For a general ω, we can write $\omega = u\pi_Q^m \,\mathrm{d}\pi_Q$ where u is a unit in $\mathcal{O}_{V,P}$ as $\Omega_{W,Q}$ is a free module generated by $\mathrm{d}\pi_Q$. Then, we can apply $\varphi^\star$ and proceed as in the first part. □

Theorem (Riemann–Hurwitz). Let $\varphi \colon V \to W$ be as above. Let $n = \deg\varphi$, $n \neq 0$. Then

$$2g(V) - 2 = n(2g(W) - 2) + \sum_{P\in V} (e_P - 1)$$

where e_P is the ramification of φ at P.

Note that the correction term $\sum_{P\in V}(e_P - 1)$ is nonnegative.

Proof. Let $\omega \in \Omega_W$ be nonzero. Then, by the Riemann–Roch theorem, and the previous lemma,

$$\begin{aligned}
2g(V) - 2 &= \deg(\operatorname{div}(\varphi^\star\omega)) \\
&= \sum_{P\in V} v_P(\varphi^\star\omega) \\
&= \sum_{Q\in W} \sum_{P\in\varphi^{-1}(Q)} v_P(\varphi^\star\omega) \\
&= \sum_{Q\in W} \sum_{P\in\varphi^{-1}(Q)} (e_P v_Q(\omega) + e_P - 1) \\
&= \sum_{Q\in W} \left(n v_Q(\omega) + \sum_{P\in\varphi^{-1}(Q)} (e_P - 1) \right) \\
&= n \deg(\operatorname{div}(\omega)) + \sum_{P\in V} (e_P - 1) \\
&= n(2g(W) - 2) + \sum_{P\in V} (e_P - 1)
\end{aligned}$$

□

Corollary. Let V, W be curves with $g(V) < g(W)$. Then any rational map $V \dashrightarrow W$ is constant.

Proof. Any rational map of this form is a morphism, then apply the Riemann–Hurwitz theorem. □

For example, there is no map $\mathbb{P}^1 \to V$ for $g(V) \geq 1$.

6.5. Equations for curves using Riemann–Roch

Let $V \subseteq \mathbb{P}^n$ be a curve not contained in any hyperplane; this can be done without loss of generality by iteratively reducing n. Let $D = \operatorname{div}(X_0)$ be the hyperplane section. Let $G \in \mathbb{C}[\mathbf{X}]$ be a homogeneous linear polynomial. Then $f = \frac{G}{X_0} \in \mathbb{C}(V)^\star$. Observe that $\operatorname{div} f + D = \operatorname{div} G$ is effective. Hence $f \in L(D)$.

We thus obtain an injective linear map from the space of linear homogeneous polynomials in $\mathbb{C}[\mathbf{X}]$ into $L(D)$ defined by $G \mapsto \frac{G}{X_0}$. This is injective because V is not contained inside a hyperplane. We make the following observations.

(i) For any point $P \in V$, there exist linear homogeneous polynomials F, G such that $F(P) \neq 0$ and $G(P) = 0$.

(ii) If P is a smooth point and L is the tangent line in $\mathbb{P}^n$, we can find a linear homogeneous polynomial F such that $F(P) = 0$ but F does not vanish on all of L.

Under this injection, we obtain the following condition. We say that a divisor D on V satisfies condition $(\star)$ if for every $P, Q \in V$ not necessarily distinct, we have $\ell(D - P - Q) = \ell(D) - 2$.

Definition. Let V be a curve, and let D a divisor with $\ell(D) = n + 1 \geq 2$. Let $\{f_0, \dots, f_n\}$ be a basis for $L(D)$. The *morphism associated to D* is $\varphi_D : V \to \mathbb{P}^n$ given by $(f_0 : \dots : f_n)$.

We say that φ_D is an *embedding* if it is an isomorphism onto its image.

Theorem. The morphism φ_D associated to D is an embedding if and only if D satisfies condition $(\star)$.

The proof is omitted.

Corollary. Suppose D is a divisor with $\deg D > 2g$. Then φ_D is an embedding.

Proof. By Riemann–Roch, D satisfies $(\star)$. □

Corollary. Every curve of genus g can be embedded in $\mathbb{P}^m$ for some m depending only on g.

Proof. If $g \geq 3$, take $[D] = 2K_V$. If $g = 2$, take $[D] = 3K_V$. If $g = 1$, take $[D] = 3[P_0]$ for some $P_0 \in V$. In any case, $\deg D > 2g$. □

Definition. A curve V of genus $g(V) \geq 2$ is called *hyperelliptic* if there exists a degree 2 morphism $V \to \mathbb{P}^1$.

The following theorem is on the last example sheet.

Theorem. A curve of genus g is hyperelliptic if and only if there exists a divisor D such that $\deg D = 2$ and $\ell(D) = 2$.

Theorem. Let V be a curve of genus $g(V) \geq 2$ that is not hyperelliptic. Then, the morphism $\varphi_{K_V} : V \to \mathbb{P}^{g-1}$ is an embedding.

Proof. Suppose that φ_K is not an embedding. Then K violates $(\star)$, so there exist points $P, Q \in V$ such that $\ell(K - P - Q) \geq g - 1$. Then by Riemann–Roch, $D = P + Q$ has $\ell(D) \geq 2$. But this is the maximal value by the above inequalities, so the result follows. □

X. Logic and Set Theory

Lectured in Lent 2023 by Prof. I. B. Leader

Mathematics is the study of logical systems, and proving true statements about them. In this course, we make precise the notion of a proof, and what it means for a logical sentence to be true. This allows us to reason about truth mathematically rather then philosophically. One important result, the completeness theorem, states that a sentence is true exactly when it has a proof. This assures us that proofs are a sensible way of showing that a statement is true, and shows us that if a statement is false there must be a counterexample.

A major application of our theory of logic is set theory. With it, we can formalise the intuitive notion of a set into a concrete mathematical object that can be studied in its own right. We can prove results about sets and set theory itself without worrying about circular logic.

To learn about the structure of the universe of sets, we will study ordinals and cardinals, which are different kinds of transfinite number. Ordinals measure discrete processes that are allowed to continue past infinity. They have rich structure, and are used to prove important and far-reaching results, such as Zorn's lemma. Cardinals measure the sizes of sets. Both ordinals and cardinals have their own arithmetic, which allow us to reason about various kinds of composition of sets and orders.

Contents

1. Propositional logic

1.1. Languages

Let P be a set of *primitive propositions*. Unless otherwise stated, we let $P = \{p_1, p_2, \dots\}$. The *language* $L = L(P)$ is defined inductively by

(i) if $p \in P$, then $p \in L$;

(ii) $\bot \in L$, where the symbol $\bot$ is read 'false';

(iii) if $p, q \in L$, then $(p \Rightarrow q) \in L$.

Example. $((p_1 \Rightarrow p_2) \Rightarrow (p_1 \Rightarrow p_3)) \in L$. $(p_4 \Rightarrow \bot) \in L$.

Remark. Note that the elements of L, called propositions, are just strings of symbols from the alphabet $\{(,), \Rightarrow, \bot, p_1, p_2, \dots\}$. Brackets are only given for clarity; we omit those that are unnecessary, and may use other types of brackets such as square brackets.

Note that the phrase 'L is defined inductively' means more precisely the following. Let $L_1 = P \cup \{\bot\}$, and define $L_{n+1} = L_n \cup \{(p \Rightarrow q) \mid p, q \in L_n\}$. We set $L = \bigcup_{n=1}^{\infty} L_n$. Note that the introduction rules for the language are injective and have disjoint ranges, so there is exactly one way in which any element of the language can be constructed using rules (i) to (iii).

We can now introduce the abbreviations $\neg, \wedge, \vee$ defined by

$$\neg p = (p \Rightarrow \bot); \quad p \vee q = \neg p \Rightarrow q; \quad p \wedge q = \neg(p \Rightarrow \neg q)$$

1.2. Semantic implication

Definition. A *valuation* is a function $v : L \to \{0, 1\}$ such that

(i) $v(\bot) = 0$;

(ii) $v(p \Rightarrow q) = 0$ if $v(p) = 1$ and $v(q) = 0$, and 1 otherwise.

Remark. On $\{0, 1\}$, we can define the constant $\bot = 0$ and the operation $\Rightarrow$ in the obvious way. Then, a valuation is precisely a mapping $L \to \{0, 1\}$ preserving all structure, so it can be considered a homomorphism.

Proposition. Let $v, v' : L \to \{0, 1\}$ be valuations that agree on the primitives p_i. Then $v = v'$. Further, any function $w : P \to \{0, 1\}$ extends to a valuation.

Remark. This is analogous to the definition of a linear map by its action on the basis vectors.

Proof. Clearly, v, v' agree on L_1, the set of elements of the language of length 1. If v, v' agree at p, q, then they agree at $p \Rightarrow q$. So by induction, v, v' agree on L_n for all n, and hence on L.

Let $v(p) = w(p)$ for all $p \in P$, and $v(\bot) = 0$ to obtain v on the set L_1. Assuming v is defined on p, q we can define it at $p \Rightarrow q$ in the obvious way. This defines v on all of L. □

Example. Let v be the valuation with $v(p_1) = v(p_3) = 1$, and $v(p_n) = 0$ for all $n \neq 1, 3$. Then, $v((p_1 \Rightarrow p_3) \Rightarrow p_2) = 0$.

Definition. A *tautology* is $t \in L$ such that $v(t) = 1$ for every valuation v. We write $\vDash t$.

Example. $p \Rightarrow (q \Rightarrow p)$.

$v(p)$	$v(q)$	$v(q \Rightarrow p)$	$v(p \Rightarrow (q \Rightarrow p))$
0	0	1	1
0	1	0	1
1	0	1	1
1	1	1	1

Since the right-hand column is always 1, $\vDash p \Rightarrow (q \Rightarrow p)$.

Example. $\neg\neg p \Rightarrow p$, which expands to $((p \Rightarrow \bot) \Rightarrow \bot) \Rightarrow p$.

$v(p)$	$v(\neg p)$	$v(\neg\neg p)$	$v(\neg\neg p \Rightarrow p)$
0	1	0	1
1	0	1	1

Hence $\vDash \neg\neg p \Rightarrow p$.

Example. $(p \Rightarrow (q \Rightarrow r)) \Rightarrow ((p \Rightarrow q) \Rightarrow (p \Rightarrow r))$. Suppose this is not a tautology. Then we have a valuation v such that $v(p \Rightarrow (q \Rightarrow r)) = 1$ and $v((p \Rightarrow q) \Rightarrow (p \Rightarrow r)) = 0$. Hence, $v(p \Rightarrow q) = 1, v(p \Rightarrow r) = 0$, so $v(p) = 1, v(r) = 0$, giving $v(q) = 1$, but then $v(p \Rightarrow (q \Rightarrow r)) = 0$ contradicting the assumption.

Definition. Let $S \subseteq L$ and $t \in L$. We say S *entails* or *semantically implies* t, written $S \vDash t$, if $v(t) = 1$ whenever $v(s) = 1$ for all $s \in S$.

Example. Let $S = \{p \Rightarrow q, q \Rightarrow r\}$, and let $t = p \Rightarrow r$. Suppose $S \not\vDash t$, so there is a valuation v such that $v(p \Rightarrow q) = 1, v(q \Rightarrow r) = 1, v(p \Rightarrow r) = 0$. Then $v(p) = 1, v(r) = 0$, so $v(q) = 1$ and $v(q) = 0$.

Definition. We say that v is a *model* of S in L if $v(s) = 1$ for all $s \in S$.

Thus, $S \vDash t$ is the statement that every model of S is also a model of t.

Remark. The notation $\vDash t$ is equivalent to $\varnothing \vDash t$.

1.3. Syntactic implication

For a notion of proof, we require a system of axioms and deduction rules. As axioms, we take (for any $p, q, r \in L$),

(i) $p \Rightarrow (q \Rightarrow p)$;

(ii) $(p \Rightarrow (q \Rightarrow r)) \Rightarrow ((p \Rightarrow q) \Rightarrow (p \Rightarrow r))$;

(iii) $((p \Rightarrow \bot) \Rightarrow \bot) \Rightarrow p$.

Remark. Sometimes, these three axioms are considered axiom *schemes*, since they are really a different axiom for each $p, q, r \in L$. These are all tautologies.

For deduction rules, we will have only the rule *modus ponens*, that from p and $p \Rightarrow q$ one can deduce q.

Definition. Let $S \subseteq L$, $t \in L$. We say S *proves* or *syntactically implies* t, written $S \vdash t$, if there exists a sequence $t_1, \dots, t_n = t$ in L such that every t_i is either

(i) an axiom;

(ii) an element of S; or

(iii) q, where $t_j = p$ and $t_k = p \Rightarrow q$ where $j, k < i$.

We say that S is the set of *premises* or *hypotheses*, and t is the *conclusion*.

Example. We will show $\{p \Rightarrow q, q \Rightarrow r\} \vdash p \Rightarrow r$.

1. $q \Rightarrow r$ (hypothesis)
2. $(q \Rightarrow r) \Rightarrow (p \Rightarrow (q \Rightarrow r))$ (axiom 1)
3. $p \Rightarrow (q \Rightarrow r)$ (modus ponens on lines 1, 2)
4. $(p \Rightarrow (q \Rightarrow r)) \Rightarrow ((p \Rightarrow q) \Rightarrow (p \Rightarrow r))$ (axiom 2)
5. $(p \Rightarrow q) \Rightarrow (p \Rightarrow r)$ (modus ponens on lines 3, 4)
6. $p \Rightarrow q$ (hypothesis)
7. $p \Rightarrow r$ (modus ponens on lines 5, 6)

Definition. If $\varnothing \vdash t$, we say t is a *theorem*, written $\vdash t$.

Example. $\vdash p \Rightarrow p$.

1. $(p \Rightarrow ((p \Rightarrow p) \Rightarrow p)) \Rightarrow ((p \Rightarrow (p \Rightarrow p)) \Rightarrow (p \Rightarrow p))$ (axiom 2)
2. $p \Rightarrow ((p \Rightarrow p) \Rightarrow p)$ (axiom 1)
3. $(p \Rightarrow (p \Rightarrow p)) \Rightarrow (p \Rightarrow p)$ (modus ponens on lines 1, 2)
4. $p \Rightarrow (p \Rightarrow p)$ (axiom 1)
5. $p \Rightarrow p$ (modus ponens on lines 3, 4)

1.4. Deduction theorem

Theorem. Let $S \subseteq L$, and $p, q \in L$. Then $S \vdash (p \Rightarrow q)$ if and only if $S \cup \{p\} \vdash q$.

Intuitively, provability corresponds to the implication connective in L.

Proof. For the forward direction, given a proof of $p \Rightarrow q$ from S, add the line p by hypothesis and deduce q from modus ponens, to obtain a proof of q from $S \cup \{p\}$.

Conversely, suppose we have a proof of q from $S \cup \{p\}$. Let $t_1, \dots, t_n$ be the lines of the proof. We will prove that $S \vdash (p \Rightarrow t_i)$ for all i.

- If t_i is an axiom, we write t_i (axiom); $t_i \Rightarrow (p \Rightarrow t_i)$ (axiom 1); $p \Rightarrow t_i$ (modus ponens).
- If $t_i \in S$, we write t_i (hypothesis); $t_i \Rightarrow (p \Rightarrow t_i)$ (axiom 1); $p \Rightarrow t_i$ (modus ponens).
- If $t_i = p$, we write the proof of $\vdash p \Rightarrow p$ given above.
- Suppose t_i is obtained by modus ponens from t_j and $t_k = t_j \Rightarrow t_i$. We may assume by induction that $S \vdash p \Rightarrow t_k$ and $S \vdash p \Rightarrow (t_j \Rightarrow t_i)$. We write

  1. $(p \Rightarrow (t_j \Rightarrow t_i)) \Rightarrow ((p \Rightarrow t_j) \Rightarrow (p \Rightarrow t_i))$ (axiom 2)
  2. $(p \Rightarrow t_j) \Rightarrow (p \Rightarrow t_i)$ (modus ponens)
  3. $p \Rightarrow t_i$ (modus ponens)

  giving $S \vdash p \Rightarrow t_i$.

□

Example. Consider $\{p \Rightarrow q, q \Rightarrow r\} \vdash p \Rightarrow r$. By the deduction theorem, it suffices to prove $\{p \Rightarrow q, q \Rightarrow r, p\} \vdash r$, which is obtained easily from modus ponens.

1.5. Soundness

We aim to show $S \vDash t$ if and only if $S \vdash t$. The direction $S \vdash t$ implies $S \vDash t$ is called *soundness*, which is a way of verifying that our axioms and deduction rule make sense. The direction $S \vDash t$ implies $S \vdash t$ is called *adequacy*, which states that our axioms are powerful enough to deduce everything that is (semantically) true.

Proposition. Let $S \subseteq L$ and $t \in L$. Then $S \vdash t$ implies $S \vDash t$.

Proof. We have a proof $t_1, \dots, t_n$ of t from S. We aim to show that any model of S is also a model of t, so if v is a valuation that maps every element of S to 1, then $v(t) = 1$. We show this by induction on the length of the proof. $v(p) = 1$ for each axiom p and for each $p \in S$. Further, $v(t_i) = 1, v(t_i \Rightarrow t_j) = 1$, then $v(t_j) = 1$. Therefore, $v(t_i) = 1$ for all i. □

1.6. Adequacy

Consider the case of adequacy where $t = \bot$. If our axioms are adequate, $S \vDash \bot$ implies $S \vdash \bot$, so $S \nvDash \bot$. We say S is *consistent* if $S \nvdash \bot$. Therefore, in an adequate system, if S has no models then S is inconsistent; equivalently, if S is consistent then it has a model.

In fact, the statement that consistent axiom sets have a model implies adequacy in general. Indeed, if $S \vDash t$, then $S \cup \{\neg t\}$ has no models, and so it is inconsistent by assumption. Then $S \cup \{\neg t\} \vdash \bot$, so $S \vdash \neg t \Rightarrow \bot$ by the deduction theorem, giving $S \vdash t$ by axiom 3.

We aim to construct a model of S given that S is consistent. Intuitively, we want to write

$$v(t) = \begin{cases} 1 & t \in S \\ 0 & t \notin S \end{cases}$$

but this does not work on the set $S = \{p_1, p_1 \Rightarrow p_2\}$ as it would evaluate p_2 to false.

We say a set $S \subseteq L$ is *deductively closed* if $p \in S$ whenever $S \vdash p$. Any set S has a *deductive closure*, which is the (deductively closed) set of statements $\{t \in L \mid S \vdash t\}$ that S proves. If S is consistent, then the deductive closure is also consistent. Computing the deductive closure before the valuation solves the problem for $S = \{p_1, p_1 \Rightarrow p_2\}$. However, if a primitive proposition p is not in S, but $\neg p$ is also not in S, this technique still does not work, as it would assign false to both p and $\neg p$.

Theorem (model existence lemma)**.** Every consistent set $S \subseteq L$ has a model.

Proof. First, we claim that for any consistent $S \subseteq L$ and proposition $p \in L$, either $S \cup \{p\}$ is consistent or $S \cup \{\neg p\}$ is consistent. If this were not the case, then $S \cup \{p\} \vdash \bot$, and also $S \cup \{\neg p\} \vdash \bot$. By the deduction theorem, $S \vdash p \Rightarrow \bot$ and $S \vdash (\neg p) \Rightarrow \bot$. But then $S \vdash \neg p$ and $S \vdash \neg\neg p$, so $S \vdash \bot$ contradicting consistency of S.

Now, L is a countable set as each L_n is countable, so we can enumerate L as $t_1, t_2, \dots$. Let $S_0 = S$, and define $S_1 = S_0 \cup \{t_1\}$ or $S_1 = S_0 \cup \{\neg t_1\}$, chosen such that S_1 is consistent. Continuing inductively, define $\overline{S} = \bigcup_{i \in \mathbb{N}} S_i$. Then, for all $t \in L$, either $t \in \overline{S}$ or $\neg t \in \overline{S}$. Note that $\overline{S}$ is consistent; indeed, if $\overline{S} \vdash \bot$, then this proof uses hypotheses only in S_n for some n, but then $S_n \vdash \bot$ contradicting consistency of S_n. Note also that $\overline{S}$ is deductively closed, so if $\overline{S} \vdash p$, we must have $p \in \overline{S}$; otherwise, $\neg p \in \overline{S}$ so $\overline{S} \vdash \neg p$, giving $\overline{S} \vdash \bot$, contradicting consistency of $\overline{S}$. Now, define the function

$$v(t) = \begin{cases} 1 & t \in \overline{S} \\ 0 & t \notin \overline{S} \end{cases}$$

We show that v is a valuation, then the proof is complete as $v(s) = 1$ for all $s \in S$. Since $\overline{S}$ is consistent, $\bot \notin \overline{S}$, so $v(\bot) = 0$.

Suppose $v(p) = 1, v(q) = 0$. Then $p \in \overline{S}$ and $q \notin \overline{S}$, and we want to show $(p \Rightarrow q) \notin \overline{S}$. If this were not the case, we would have $(p \Rightarrow q) \in \overline{S}$ and $p \in \overline{S}$, so $q \in \overline{S}$ as $\overline{S}$ is deductively closed.

Now suppose $v(q) = 1$, so $q \in \overline{S}$, and we need to show $(p \Rightarrow q) \in \overline{S}$. Then $\overline{S} \vdash q$, and by axiom 1, $\overline{S} \vdash q \Rightarrow (p \Rightarrow q)$. Therefore, as $\overline{S}$ is deductively closed, $(p \Rightarrow q) \in \overline{S}$.

Finally, suppose $v(p) = 0$, so $p \notin \overline{S}$, and we want to show $(p \Rightarrow q) \in \overline{S}$. We know that $\neg p \in \overline{S}$, so it suffices to show that $p \Rightarrow \bot \vdash p \Rightarrow q$. By the deduction theorem, this is equivalent to proving $\{p, p \Rightarrow \bot\} \vdash q$, or equivalently, $\bot \vdash q$. But by axiom 1, $\bot \Rightarrow (\neg q \Rightarrow \bot)$ where $(\neg q \Rightarrow \bot) = \neg\neg q$, so the proof is complete by axiom 3. □

Remark. We used the fact that P was a countable set in order to show that L was countable. The result does in fact hold if P is uncountable, but requires more tools to prove. Some sources call this theorem the 'completeness theorem'.

Corollary (adequacy)**.** Let $S \subseteq L$ and let $t \in L$, such that $S \vDash t$. Then $S \vdash t$.

Proof. Follows from the remarks before the model existence lemma. □

1.7. Completeness

Theorem (completeness theorem for propositional logic)**.** Let $S \subseteq L$ and $t \in L$. Then $S \vDash t$ if and only if $S \vdash t$.

Proof. Follows from soundness and adequacy. □

Theorem (compactness theorem)**.** Let $S \subseteq L$ and $t \in L$ with $S \vDash t$. Then there exists a finite subset $S' \subseteq S$ such that $S' \vDash t$.

Proof. Trivial after applying the completeness theorem, since proofs depend on only finitely many hypotheses in S. □

Corollary (compactness theorem, equivalent form)**.** Let $S \subseteq L$. Then if every finite subset $S' \subseteq S$ has a model, then S has a model.

Proof. Let $t = \bot$ in the compactness theorem. Then, if $S \vDash \bot$, some finite $S' \subseteq S$ has $S' \vDash \bot$. But this is not true by assumption, so there is a model for S. □

Remark. This corollary is equivalent to the more general compactness theorem, since the assertion that $S \vDash t$ is equivalent to the statement that $S \cup \{\neg t\}$ has no model, and $S' \vDash t$ is equivalent to the statement that $S' \cup \{\neg t\}$ has no model.

Theorem (decidability theorem)**.** Let $S \subseteq L$ and $t \in L$. Then, there is an algorithm to decide (in finite time) if $S \vdash t$.

Proof. Trivial after replacing $\vdash$ with $\vDash$, by drawing the relevant truth tables. □

2. Well-orderings

2.1. Definition

Definition. A *total order* or *linear order* is a pair $(X, <)$ where X is a set, and $<$ is a relation on X such that

- (irreflexivity) for all $x \in X$, $x \not< x$;
- (transitivity) for all $x, y, z \in X$, $x < y$ and $y < z$ implies $x < z$;
- (trichotomy) for all $x, y \in X$, either $x < y$, $y < x$, or $x = y$.

We use the obvious notation $x > y$ to denote $y < x$. In terms of the $\leq$ relation, we can equivalently write the axioms of a total order as

- (reflexivity) for all $x \in X$, $x \leq x$;
- (transitivity) for all $x, y, z \in X$, $x \leq y$ and $y \leq z$ implies $x \leq z$;
- (antisymmetry) for all $x, y \in X$, if $x \leq y$ and $y \leq x$ then $x = y$.
- (trichotomy, or totality) for all $x, y \in X$, either $x \leq y$ or $y \leq x$.

Example. (i) $(\mathbb{N}, \leq)$ is a total order.

(ii) $(\mathbb{Q}, \leq)$ is a total order.

(iii) $(\mathbb{R}, \leq)$ is a total order.

(iv) $(\mathbb{N}^+, |)$ is not a total order, where $|$ is the divides relation, since 2 and 3 are not related.

(v) $(\mathcal{P}(S), \subseteq)$ is not a total order if $|S| > 1$, since it fails trichotomy.

Definition. A total order $(X, <)$ is a *well-ordering* if every nonempty subset $S \subseteq X$ has a least element.

$$\forall S \subseteq X,\ S \neq \varnothing \implies \exists x \in S,\ \forall y \in S,\ x \leq y$$

Example. (i) $(\mathbb{N}, <)$ is a well-ordering.

(ii) $(\mathbb{Z}, <)$ is not a well-ordering, since $\mathbb{Z}$ has no least element.

(iii) $(\mathbb{Q}, <)$ is not a well-ordering.

(iv) $(\mathbb{R}, <)$ is not a well-ordering.

(v) $[0,1] \subset \mathbb{R}$ with the usual order is not a well-ordering, since $(0,1]$ has no least element.

(vi) $\left\{\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \dots\right\} \subset \mathbb{R}$ with the usual order is a well-ordering.

(vii) $\left\{\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \dots\right\} \cup \{1\}$ with the usual order is also a well-ordering.

(viii) $\left\{\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \dots\right\} \cup \{2\}$ with the usual order is another example.

(ix) $\left\{\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \dots\right\} \cup \left\{1 + \frac{1}{2}, 1 + \frac{2}{3}, 1 + \frac{3}{4}, \dots\right\}$ is another example.

Remark. Let $(X, <)$ be a total order. $(X, <)$ is a well-ordering if and only if there is no infinite decreasing sequence $x_1 > x_2 > \dots$. Indeed, if $(X, <)$ is a well-ordering, then the set $\{x_1, x_2, \dots\}$ has no minimal element, contradicting the assumption. Conversely, if $S \subseteq X$ has no minimal element, then we can construct an infinite decreasing sequence by arbitrarily choosing points $x_1 > x_2 > \dots$ in S, which exists as S has no minimal element.

Definition. Total orders X, Y are *isomorphic* if there is a bijection f between X and Y that preserves $<$: $x < y$ if and only if $f(x) < f(y)$.

Examples (i) and (vi) are isomorphic, and (vii) and (viii) are isomorphic. Examples (i) and (vii) are not isomorphic, since example (vii) has a greatest element and (i) does not.

Proposition (proof by induction). Let X be a well-ordered set, and let $S \subseteq X$ such that

$$\forall x \in S, (\forall y < x, y \in S) \implies x \in S$$

Then $S = X$.

Remark. Equivalently, if $p(x)$ is a property such that if $p(y)$ is true for all $y < x$ then $p(x)$, then $p(x)$ holds for all x.

Proof. Suppose $S \neq X$. Then $X \setminus S$ is nonempty, and therefore has a least element x. But all elements $y < x$ lie in S, and so by the property of S, we must have $x \in S$, contradicting the assumption. □

Proposition. Let X, Y be isomorphic well-orderings. Then there is exactly one isomorphism between X and Y.

Note that this does not hold for general total orderings, such as $\mathbb{Q}$ to itself or $[0, 1]$ to itself.

Proof. Let $f, g : X \to Y$ be isomorphisms. We show that $f(x) = g(x)$ for all x by induction on x. Suppose $f(y) = g(y)$ for all $y < x$. We must have that $f(x) = a$, where a is the least element of $Y \setminus \{f(y) \mid y < x\}$. Indeed, if not, we have $f(x') = a$ for some $x' > x$ by bijectivity, contradicting the order-preserving property. Note that the set $Y \setminus \{f(x) \mid y < x\}$ is nonempty as it contains $f(x)$. So $f(x) = a = g(x)$, as required. □

2.2. Initial segments

Definition. A subset I of a totally ordered set X is an *initial segment* if $x \in I$ implies $y \in I$ for all $y < x$.

Example. In any total ordering X and element $x \in X$, the set $\{y \mid y < x\}$ is an initial segment. Not every initial segment is of this form, for instance $\{x \mid x \leq 3\}$ in $\mathbb{R}$, or $\left\{x \mid x > 0, x^2 < 2\right\}$ in $\mathbb{Q}$.

In a well-ordering, every proper initial segment $I \neq X$ is of this form. Indeed, $I = \{y \mid y < x\}$ where x is the least element of $X \setminus I$: $y \in I$ implies $y < x$, otherwise $y = x$ or $x < y$, giving the contradiction $x \in I$; and conversely, $y < x$ implies $y \in I$, otherwise y is a smaller element of $X \setminus I$.

Theorem (definition by recursion)**.** Let X be a well-ordering and Y be any set. Let $G : \mathcal{P}(X \times Y) \to Y$ be a rule that assigns a point in Y given a definition of the function 'so far', represented as a set of ordered pairs. Then there exists a function $f : X \to Y$ such that $f(x) = G\big(f|_{I_x}\big)$, and such a function is unique.

Remark. In defining $f(x)$, we may use the value of $f(y)$ for all $y < x$.

Proof. We say that h is an *attempt* to mean that $h : I \to Y$ where I is some initial segment of X, and for all $x \in I$ we have that $h(x) = G\big(h|_{I_x}\big)$. Note that if h, h' are attempts both defined at x, then $h(x) = h'(x)$ by induction on x.

Also, for all x, there exists an attempt defined at x, by induction on x. Indeed, by induction we can assume there exists an attempt h_y defined at y for all $y < x$, and then we can define h to be the union of the h_y. This is an attempt with domain I_x, so the attempt $h' = h \cup \{(x, G(h))\}$ is an attempt defined at x. Therefore, there is an attempt defined at each x, so we can define the function $f : X \to Y$ by setting $f(x)$ to be the value of $h(x)$ where h is some attempt defined at x.

For uniqueness, we apply induction on x. If f, f' agree below x, then they must agree at x since $f(x) = G\big(f|_{I_x}\big) = G\big(f'|_{I_x}\big) = f'(x)$. □

Proposition (subset collapse)**.** Any subset Y of a well-ordering X is isomorphic to a unique initial segment of X.

This is not true for general total orderings, such as $\{1, 2, 3\} \subset \mathbb{Z}$, or $\mathbb{Q}$ in $\mathbb{R}$.

Proof. If f is some such isomorphism, we must have that $f(x)$ is the least element of X not of the form $f(y)$ for $y < x$. We define f in this way by recursion, and this is an isomorphism as required. Note that this is always well-defined as $f(y) \leq y$, so there is always some element of X (namely, x) not of the form $f(y)$ for $y < x$. Uniqueness follows by induction. □

Remark. X itself cannot be isomorphic to a proper initial segment by uniqueness as it is isomorphic to itself.

2.3. Relating well-orderings

Definition. For well-orderings X, Y, we will write $X \leq Y$ if X is isomorphic to an initial segment of Y.

$X \leq Y$ if and only if X is isomorphic to some subset of Y.

Example. $\mathbb{N} \leq \left\{\frac{1}{2}, \frac{2}{3}, \dots\right\}$.

Proposition. Let X, Y be well-orderings. Then either $X \leq Y$ or $Y \leq X$.

Proof. By recursion we define the function $f : X \to Y$ by letting $f(x)$ be the least element of Y not of the form $f(y)$ for all $y < x$. If a least element of this form always exists, this is a well-defined isomorphism from X to an initial segment of Y as required. Suppose that $Y \setminus \{f(y) \mid y < x\}$ is empty, so $\{f(y) \mid y < x\} = Y$. Then Y is isomorphic to $I_x \subseteq X$, so $Y \leq X$. □

Proposition. Let X, Y be well-orderings, and suppose $X \leq Y$ and $Y \leq X$. Then X is isomorphic to Y.

Proof. Let $f : X \to Y$ and $g : Y \to X$ be isomorphisms to initial segments. Then $g \circ f$ is an isomorphism from X to some initial segment of X, as an initial segment of an initial segment is an initial segment. So by uniqueness, $g \circ f$ is the identity map on X. Similarly, $f \circ g$ is the identity on Y, so f and g are inverses. □

2.4. Constructing larger well-orderings

Definition. For well-orderings X, Y, we write $X < Y$ if $X \leq Y$ and X is not isomorphic to Y.

Equivalently, $X < Y$ if X is isomorphic to a proper initial segment of Y.

Let X be a well-ordering, and let $x \notin X$. Construct the well-ordering on $X \cup \{x\}$ by setting $y < x$ for all $y \in X$. This well-ordering is strictly greater than X, since X is isomorphic to a proper initial segment. This is called the *successor* of X, written X^+.

For well-orderings $(X, <_X), (Y, <_Y)$, we say that $(Y, <_Y)$ *extends* $(X, <_X)$ if $X \subseteq Y$, $<_Y |_X = <_X$, and X is an initial segment of Y. We say that well-orderings X_i for $i \in I$ are *nested* if for all $i, j \in I$, either X_i extends X_j or X_j extends X_i.

Proposition. Let X_i for $i \in I$ be a nested set of well-orderings. Then, there exists a well-ordering X such that $X_i \leq X$ for all $i \in I$.

Proof. Let $X = \bigcup_{i \in I} X_i$ with ordering $<_X = \bigcup_{i \in I} <_i$. Then, as the X_i are nested, each X_i is an initial segment of X. We show that this is a well-ordering. Let $S \subseteq X$ be a nonempty set. Then $S \cap X_i \neq \varnothing$ for some $i \in I$. Let x be the least element of $S \cap X_i$. Thus, x is the least element of S, as X_i is an initial segment of X. □

Remark. The proposition holds without the nestedness assumption.

2.5. Ordinals

Definition. An *ordinal* is a well-ordered set, where we regard two ordinals as equal if they are isomorphic.

Remark. We cannot construct ordinals as equivalence classes of well-orderings, due to Russell's paradox. Later, we will see a different construction that deals with this problem.

Definition. Let X be a well-ordering corresponding to an ordinal α. Then, we say that X has *order type* α.

The order type of the unique well-ordering on a collection of $k \in \mathbb{N}$ points is named k. The order type of $(\mathbb{N}, <)$ is named ω.

Example. In the reals, the set $\{-2, 3, -\pi, 5\}$ has order type 4. The set $\left\{\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \dots\right\}$ has order type ω.

We will write $\alpha \leq \beta$ if $X \leq Y$ where X has order type α and Y has order type β. This does not depend on the choice of representative X or Y. We define $\alpha < \beta$ and α^+ in a similar way. Note that $\alpha \leq \beta, \beta \leq \alpha$ implies $\alpha = \beta$. Therefore, ordinals are totally ordered.

Proposition. Let α be an ordinal. Then the set of ordinals less than α form a well-ordered set of order type α.

Proof. Let X be a well-ordering with order type α. Then, the well-orderings less than X are precisely the proper initial segments of X, up to isomorphism. The initial segments of X are precisely the sets $I_x = \{y \in X \mid y < x\}$ for $x \in X$. But these are order isomorphic to X itself by mapping $I_x \mapsto x$. □

We define $I_\alpha = \{\beta < \alpha\}$, which is a well-ordered set of order type α. This is often a convenient representative to choose for an ordinal.

Proposition. Every nonempty set S of ordinals has a least element.

Proof. Let $\alpha \in S$. Suppose α is not the least element of S. Then $S \cap I_\alpha$ is nonempty. But I_α is well-ordered, so $S \cap I_\alpha$ has a minimal element as required. □

Theorem (Burali-Forti paradox)**.** The ordinals do not form a set.

Proof. Suppose X is the set of all ordinals. Then X is a well-ordered set, so it has an order type α. Then X is isomorphic to I_α, which is a proper initial segment of X. □

Remark. Given a set $S = \{\alpha_i : i \in I\}$ of ordinals, there exists an upper bound α for S, so $\alpha_i \leq \alpha$ for all $i \in I$, by considering the nested family of well-orderings I_{α_i}. Hence, by the previous proposition, there exists a least upper bound, as I_α is a set. We write $\alpha = \sup S$.

Example. $\sup\{2, 4, 6, \dots\} = \omega$.

Remark. If we represent ordinals by sets of smaller ordinals, $\sup S = \bigcup_{\alpha \in S} \alpha$.

2.6. Some ordinals

$$0, 1, 2, 3, \dots, \omega$$

Write $\alpha + 1$ for the successor α^+ of α.

$$\omega + 1, \omega + 2, \omega + 3, \dots, \omega + \omega = \omega \cdot 2$$

where $\omega + \omega = \omega \cdot 2$ is defined by $\sup\{\omega, \omega + 1, \omega + 2, \dots\}$.

$$\omega \cdot 2 + 1, \omega \cdot 2 + 2, \dots, \omega \cdot 3, \omega \cdot 4, \omega \cdot 5, \dots, \omega \cdot \omega = \omega^2$$

where we define $\omega \cdot \omega = \sup\{\omega \cdot 2, \omega \cdot 3, \dots\}$.

$$\omega^2 + 1, \omega^2 + 2, \dots, \omega^2 + \omega, \dots, \omega^2 + \omega \cdot 2, \dots, \omega^2 + \omega^2 = \omega^2 \cdot 2$$

Continue in the same way.

$$\omega^2 \cdot 3, \omega^2 \cdot 4, \dots, \omega^3$$

where $\omega^3 = \sup\left\{\omega^2 \cdot 2, \omega^2 \cdot 3, \dots\right\}$.

$$\omega^3 + \omega^2 \cdot 7 + \omega \cdot 4 + 13, \dots, \omega^4, \omega^5, \dots, \omega^\omega$$

where $\omega^\omega = \sup\left\{\omega, \omega^2, \omega^3, \dots\right\}$.

$$\omega^\omega \cdot 2, \omega^\omega \cdot 3, \dots, \omega^\omega \cdot \omega = \omega^{\omega+1}$$

$$\omega^{\omega+2}, \dots, \omega^{\omega\cdot 2}, \omega^{\omega\cdot 3}, \dots, \omega^{\omega^2}, \dots, \omega^{\omega^3}, \dots, \omega^{\omega^\omega}, \dots, \omega^{\omega^{\omega^\omega}}, \dots, \omega^{\omega^{\omega^{\cdots}}} = \varepsilon_0$$

where $\varepsilon_0 = \sup\left\{\omega, \omega^\omega, \omega^{\omega^\omega}, \dots\right\}$.

$$\varepsilon_0 + 1, \varepsilon_0 + \omega, \varepsilon_0 + \varepsilon_0 = \varepsilon_0 \cdot 2, \dots, \varepsilon_0^2, \varepsilon_0^3, \dots, \varepsilon_0^{\varepsilon_0}$$

where $\varepsilon_0^{\varepsilon_0} = \sup\left\{\varepsilon_0^\omega, \varepsilon_0^{\omega^\omega}, \dots\right\}$.

$$\varepsilon_0^{\varepsilon_0^{\varepsilon_0^{\cdots}}} = \varepsilon_1$$

All of these ordinals are countable, as each operation only takes a countable union of countable sets.

2.7. Uncountable ordinals

Theorem. There exists an uncountable ordinal.

Remark. The reals cannot be explicitly well-ordered.

Proof. Let $A \subseteq \mathcal{P}(\omega \times \omega)$ be the set of well-orderings of subsets of $\mathbb{N}$. Let B be the set of order types of A. Then B is the set of all countable ordinals. Let $\omega_1 = \sup B$. ω_1 is uncountable, and in particular, the least uncountable ordinal. Indeed, if it were countable, it would be the greatest element of B, but $\omega_1 + 1$ would also lie in B. □

Remark. Without introducing A, it would be difficult to show that B was in fact a set.

Remark. Another ending to the proof above is as follows. B cannot be the set of all ordinals, since the ordinals do not form a set by the Burali-Forti paradox, so there exists an uncountable ordinal. In particular, there exists a least uncountable ordinal.

The ordinal ω_1 has a number of remarkable properties.

(i) ω_1 is uncountable, but $\{\beta \mid \beta < \alpha\}$ is countable for all $\alpha < \omega_1$.

(ii) There exists no sequence $\alpha_1, \alpha_2, \dots$ in I_{ω_1} with supremum ω_1, as it is bounded by $\sup\{\alpha_1, \alpha_2, \dots\}$, which is a countable ordinal.

Theorem (Hartogs' lemma)**.** For every set X, there exists an ordinal γ that does not inject into X.

Proof. Use the argument above from the existence of an uncountable ordinal. □

We write $\gamma(X)$ for the least ordinal that does not inject into X. For example $\gamma(\omega) = \omega_1$.

2.8. Successors and limits

Definition. We say that an ordinal α is a *successor* if there exists β such that $\alpha = \beta^+$. Otherwise, α is a *limit*.

Equivalently, an ordinal is a successor if and only if it has a greatest element. An ordinal α is a limit if and only if it has no greatest element, or equivalently, for all $\beta < \alpha$, there exists $\gamma < \alpha$ with $\gamma > \beta$, giving $\alpha = \sup\{\beta \mid \beta < \alpha\}$.

Example. 5 is a successor. $\omega + 2 = (\omega^+)^+$ is a successor. ω is a limit as it has no greatest element. 0 is a limit.

2.9. Ordinal arithmetic

Let α, β be ordinals. We define $\alpha + \beta$ by induction on β, by

- $\alpha + 0 = \alpha$;
- $\alpha + \beta^+ = (\alpha + \beta)^+$;
- $\alpha + \lambda = \sup\{\alpha + \gamma \mid \gamma < \lambda\}$ for a nonzero limit ordinal.

Example. $\omega + 1 = \omega + 0^+ = (\omega + 0)^+ = \omega^+$. $\omega + 2 = \omega + 1^+ = (\omega + 1)^+ = (\omega^+)^+$. $1 + \omega = \sup\{1 + \gamma \mid \gamma < \omega\} = \omega$. Therefore, addition is noncommutative.

Remark. As the ordinals do not form a set, we must technically define addition $\alpha + \gamma$ by induction on the set $\{\gamma \mid \gamma \leq \beta\}$. The choice of β does not change the definition of $\alpha + \gamma$ as defined for $\gamma \leq \beta$.

Proposition. Ordinal addition is associative.

Proof. Let α, β, γ be ordinals. We use induction on γ. Suppose $\alpha + (\beta + \delta) = (\alpha + \beta) + \delta$ for all $\delta < \gamma$.

First, suppose $\gamma = 0$. $\alpha + (\beta + 0) = \alpha + \beta = (\alpha + \beta) = 0$, as required. Now consider γ^+.

$$\alpha + (\beta + \gamma^+) = \alpha + (\beta + \gamma)^+ = (\alpha + (\beta + \gamma))^+ = ((\alpha + \beta) + \gamma)^+ = (\alpha + \beta) + \gamma^+$$

Finally, consider λ a nonzero limit.

$$(\alpha + \beta) + \lambda = \sup\{(\alpha + \beta) + \gamma \mid \gamma < \lambda\} = \sup\{\alpha + (\beta + \gamma) \mid \gamma < \lambda\}$$

We claim that $\beta + \lambda$ is a limit. Indeed, $\beta + \lambda = \sup\{\beta + \gamma \mid \gamma < \lambda\}$, but for every $\gamma < \lambda$ there exists $\gamma' < \lambda$ with $\gamma < \gamma'$ as λ is a limit, so $\beta + \gamma < \beta + \gamma'$. Thus, there is no greatest element in the set $\{\beta + \gamma \mid \gamma < \lambda\}$, so $\beta + \lambda$ is a limit.

Now, $\alpha + (\beta + \lambda) = \sup\{\alpha + \delta \mid \delta < \beta + \lambda\}$. So it suffices to show that

$$\sup\{\alpha + (\beta + \gamma) \mid \gamma < \lambda\} = \sup\{\alpha + \delta \mid \delta < \beta + \lambda\}$$

Certainly

$$\{\alpha + (\beta + \gamma) \mid \gamma < \lambda\} \subseteq \{\alpha + \delta \mid \delta < \beta + \lambda\}$$

as $\gamma < \lambda$ implies $\beta + \gamma < \beta + \lambda$. Further, for any $\delta < \beta + \lambda$, $\delta \leq \beta + \gamma$ for some $\gamma < \lambda$ by definition of $\beta + \lambda$. Therefore, $\alpha + \delta \leq \alpha + (\beta + \gamma)$, so each element of $\{\alpha + \delta \mid \delta < \beta + \lambda\}$ is at most some element of $\{\alpha + (\beta + \gamma) \mid \gamma < \lambda\}$. So the two suprema agree. □

Remark. We used the facts

(i) $\beta \leq \gamma \implies \alpha + \beta \leq \alpha + \gamma$, which is trivial by induction on γ;

(ii) $\beta < \gamma \implies \alpha + \beta < \alpha + \gamma$, as $\beta^+ \leq \gamma$ so $\alpha + \beta^+ \leq \alpha + \gamma$ by (i).

However, $1 < 2$ but $1 + \omega \not< 2 + \omega$.

The above is the *inductive* definition of addition; there is also a *synthetic* definition of addition. We can define $\alpha + \beta$ to be the order type of $\alpha \sqcup \beta$, where every element of α is taken to be less than every element of β.

For instance, $\omega + 1$ is the order type of ω with a point afterwards, and $1 + \omega$ is the order type of a point followed by ω, which is clearly isomorphic to ω. Associativity is clear, as $(\alpha+\beta)+\gamma$ and $\alpha + (\beta + \gamma)$ are the order type of $\alpha \sqcup \beta \sqcup \gamma$.

Proposition. The inductive and synthetic definitions of addition coincide.

Proof. We write $+'$ for synthetic addition, and aim to show $\alpha + \beta = \alpha +' \beta$. We perform induction on β.

For $\beta = 0$, $\alpha + 0 = \alpha$ and $\alpha +' 0 = \alpha$. For successors, $\alpha + \beta^+ = (\alpha + \beta)^+ = (\alpha +' \beta)^+$, which is the order type of $\alpha \sqcup \beta \sqcup \{\star\}$, which is equal to $\alpha +' \beta^+$.

Let λ be a nonzero limit. We have $\alpha + \lambda = \sup\{\alpha + \gamma \mid \gamma < \lambda\}$. But $\alpha + \gamma = \alpha +' \gamma$ for $\gamma < \lambda$, so $\alpha + \lambda = \sup\{\alpha +' \gamma \mid \gamma < \lambda\}$. As the set $\{\alpha +' \gamma \mid \gamma < \lambda\}$ is nested, it is equal to its union, which is $\alpha +' \lambda$. □

Synthetic definitions can be easier to work with if such definitions exist. However, there are many definitions that can only easily be represented inductively, and not synthetically.

We define multiplication inductively by

- $\alpha 0 = 0$;
- $\alpha\beta^+ = \alpha\beta + \alpha$;
- $\alpha\lambda = \sup\{\alpha\gamma \mid \gamma < \lambda\}$ for λ a nonzero limit.

Example. $\omega 2 = \omega 1 + \omega = \omega 0 + \omega + \omega = \omega + \omega$. Similarly, $\omega 3 = \omega + \omega + \omega$. $\omega\omega = \sup\{0, \omega 1, \omega 2, \dots\} = \{0, \omega, \omega + \omega, \dots\}$. Note that $2\omega = \sup\{0, 2, 4, \dots\} = \omega$. Multiplication is noncommutative. One can show in a similar way that multiplication is associative.

We can produce a synthetic definition of multiplication, which can be shown to coincide with the inductive definition. We define $\alpha\beta$ to be the order type of the Cartesian product $\alpha \times \beta$ where we say $(\gamma, \delta) < (\gamma', \delta')$ if $\delta < \delta'$ or $\delta = \delta'$ and $\gamma < \gamma'$. For instance, $\omega 2$ is the order type of two infinite sequences, and 2ω is the order type of a sequence of pairs.

Similar definitions can be created for exponentiation, towers, and so on. For instance, α^β can be defined by

- $\alpha^0 = 1$;
- $\alpha^{(\beta^+)} = \alpha^\beta \alpha$;
- $\alpha^\lambda = \sup\{\alpha^\gamma \mid \gamma < \lambda\}$ for λ a nonzero limit.

For example, $\omega^2 = \omega^1\omega = \omega^0\omega\omega = \omega\omega$. Further, $2^\omega = \sup\{2^0, 2^1, \dots\} = \omega$, which is countable.

3. Posets

3.1. Definitions

Definition. A *partially ordered set* or *poset* is a pair $(X, \leq)$ where X is a set, and $\leq$ is a relation on X such that

- (reflexivity) for all $x \in X$, $x \leq x$;
- (transitivity) for all $x, y, z \in X$, $x \leq y$ and $y \leq z$ implies $x \leq z$;
- (antisymmetry) for all $x, y \in X$, $x \leq y$ and $y \leq x$ implies $x = y$.

We write $x < y$ for $x \leq y$ and $x \neq y$. Alternatively, a post is a pair $(X, <)$ where X is a set, and $<$ is a relation on X such that

- (irreflexivity) for all $x \in X$, $x \not< x$;
- (transitivity) for all $x, y, z \in X$, $x < y$ and $y < z$ implies $x < z$.

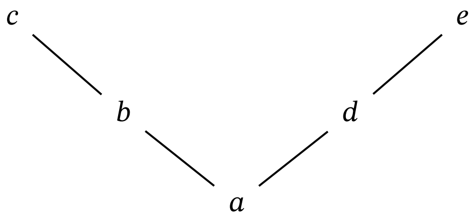
Example. (i) Any total order is a poset.

(ii) $\mathbb{N}^+$ with the divides relation is a poset.

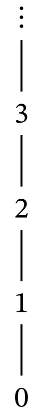
(iii) $(\mathcal{P}(S), \subseteq)$ is a poset.

(iv) $(X, \subseteq)$ is a poset where $X \subseteq \mathcal{P}(S)$, such as the set of vector subspaces of a vector space.

(v) The following diagram is also a poset, where the lines from a upwards to b denote relations $a \leq b$.

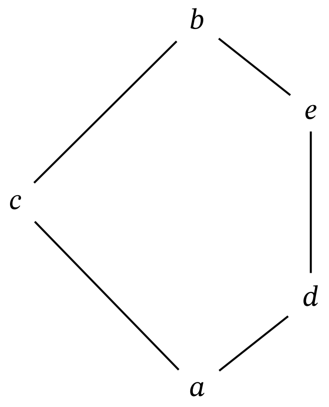


This is called a *Hasse diagram*. An upwards line from x to y is drawn if *y covers x*, so $y > x$ and no z has $y > z > x$. The natural numbers can be represented as a Hasse diagram.

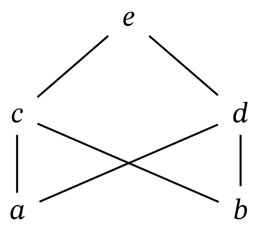


The rationals cannot, since no element covers another.

(vi) There is no notion of ‘height’ in a poset, illustrated by the following diagram.



(vii)



Definition. A subset S of a poset X is a *chain* if it is totally ordered.

Example. The powers of 2 in $(\mathbb{N}^+, |)$ is a chain.

Definition. A subset S of a poset X is an *antichain* if no two distinct elements are related.

Example. The set of primes in $(\mathbb{N}^+, |)$ is an antichain.

Definition. For $S \subseteq X$, an *upper bound* for S is an $x \in X$ such that $x \geq y$ for all $y \in S$. A *least upper bound* is an upper bound $x \in X$ for S such that for all upper bounds $y \in X$ for S, $x \leq y$.

Example. If $S = \left\{x \mid x < \sqrt{2}\right\} \subset \mathbb{R}$, 7 is an upper bound, and $\sqrt{2}$ is a least upper bound. We write $\sqrt{2} = \sup S = \bigvee S$ for the least upper bound or *join* of S.

In $\mathbb{Q}$, the set $\left\{x \mid x^2 < 2\right\}$ has 7 as an upper bound but has no least upper bound.

In example (v), $\{a, b\}$ has upper bounds b and c, so the least upper bound is b. $\{b, d\}$ has no upper bound. In example (vii), $\{a, b\}$ has upper bounds c, d, e, so does not have a least upper bound.

Definition. A poset X is *complete* if every $S \subseteq X$ has a least upper bound.

Example. $\mathbb{R}$ is not complete, as $\mathbb{Z}$ has no upper bound. $[0, 1] \subseteq \mathbb{R}$ is complete. $(0, 1) \subseteq \mathbb{R}$ is not complete, as $(0, 1)$ has no upper bound.

Example. $X = \mathcal{P}(S)$ is always complete as a poset under inclusion, with $\sup\{A_i \mid i \in I\} = \bigcup_{i \in I} A_i$.

Note that every complete poset X has a greatest element $\sup X$. A complete poset also has a least element $\sup \varnothing$. In the case $X = \mathcal{P}(S)$, $\sup X = S$ and $\sup \varnothing = \varnothing$.

Definition. Let $f : X \to Y$ be a function where X, Y are posets. We say f is *order-preserving* if $x \leq y$ implies $f(x) \leq f(y)$.

Example. The function $f : \mathbb{N} \to \mathbb{N}$ defined by $f(x) = x + 1$ is order-preserving. The function $f : [0, 1] \to [0, 1]$ defined by $x \mapsto \frac{x+1}{2}$ is order-preserving. The function $f : \mathcal{P}(S) \to \mathcal{P}(S)$ defined by $f(A) = A \cup \{i\}$ for some fixed $i \in S$ is order-preserving.

Not all order-preserving functions have a fixed point x such that $f(x) = x$, for example $f(x) = x + 1$ on $\mathbb{N}$.

Theorem (Knaster–Tarski fixed point theorem). Let X be a complete poset. Then every order-preserving $f : X \to X$ has a fixed point.

Proof. Let $E = \{x \in X \mid x \leq f(x)\}$, and let $s = \sup E$. We show that s is a fixed point for f.

First, we show $s \leq f(s)$, so $s \in E$. It suffices to show $f(s)$ is an upper bound for E, then the result holds as s is the least such upper bound. If $x \in E$, we know $x \leq s$, so $f(x) \leq f(s)$ as f is order-preserving, as required.

Now, we show $f(s) \leq s$. It suffices to show $f(s) \in E$, as s is an upper bound for E. Since $s \leq f(s)$, we have $f(s) \leq f(f(s))$, but this is precisely the fact that $f(s) \in E$. □

Corollary (Schröder–Bernstein theorem). Let $f : A \to B$ and $g : B \to A$ be injections. Then there is a bijection $A \to B$.

Proof. We seek partitions $A = P \sqcup Q$, $B = R \sqcup S$ such that $f(P) = R$ and $g(S) = Q$; then we define h to equal to f on P and g^{-1} on Q. Thus, we need a set P that is a fixed point of $\theta : \mathcal{P}(A) \to \mathcal{P}(A)$ given by $P \mapsto A \setminus g(B \setminus f(P))$. But θ is order-preserving and $\mathcal{P}(A)$ is a complete poset. So P exists by the Knaster–Tarski fixed point theorem. □

3.2. Zorn's lemma

Definition. Let X be a poset. We say that $x \in X$ is *maximal* if there is no $y \in X$ with $y > x$.

Example. In $[0,1]$, 1 is maximal. In example (v), there are two maximal elements c and e.

Note that $(\mathbb{R}, \leq)$ and $(\mathbb{N}, |)$ have no maximal elements, and they both have a chain with no upper bound, such as $\mathbb{N} \subset \mathbb{R}$, and powers of two.

Theorem (Zorn's lemma)**.** Let X be a poset in which every chain has an upper bound. Then X has a maximal element.

The empty chain must have an upper bound in X, so X must be nonempty to apply Zorn's lemma. Zorn's lemma can be equivalently be stated as the following.

Theorem. Let X be a nonempty poset in which every nonempty chain has an upper bound. Then X has a maximal element.

One can view Zorn's lemma as a fixed point theorem on a function $f : X \to X$ with the property that $x \leq f(x)$.

Proof. Suppose that X has no maximal element. Then for each $x \in X$, we have $x' \in X$ and $x' > x$. For each chain C, we have an upper bound $u(C)$. Let $x \in X$ be any element, and define x_α for each $\alpha < \gamma(X)$ by recursion.

- $x_0 = x$;
- $x_{\alpha+1} = x'_\alpha$;
- $x_\lambda = u\{x_\beta \mid \beta < \lambda\}$ for λ a nonzero limit.

Note that $\{x_\beta \mid \beta < \lambda\}$ forms a chain, so it has an upper bound as required. Then, we have an injection from $\gamma(X)$ into X, contradicting the definition of $\gamma(X)$. □

Remark. Although this proof was short, it relied on the infrastructure of well-orderings, recursion, ordinals, and Hartogs' lemma.

We show that every vector space has a basis. Recall that a basis is a linearly independent spanning set; no nontrivial finite linear combination of basis elements is zero, and each element of the vector space is a finite linear combination of the basis elements. For instance, the space of real polynomials has basis $1, X, X^2, \dots$. The space of real sequences has a linearly independent set $(1,0,0,\dots), (0,1,0,\dots), \dots$, but this is not a basis as the sequence $(1,1,1,\dots)$ cannot be constructed as a finite linear combination of these vectors. In fact, there is no countable basis for this space, and no explicitly definable basis in general. $\mathbb{R}$ is a vector space over $\mathbb{Q}$. There is clearly no countable basis, and in fact no explicit basis. A basis in this case is called a *Hamel basis*.

Theorem. Every vector space V has a basis.

Proof. Let X be the set of all linearly independent subsets of V, ordered by inclusion. We seek a maximal element of X; this is clearly a basis, as any vector not in its span could be added to the set to increase the set of basis vectors. X is nonempty as $\varnothing \in X$.

We apply Zorn's lemma. Let $(A_i)_{i\in I}$ be a chain in X. We show that its union $A = \bigcup_{i\in I} A_i$ is a linearly independent set, and therefore lies in X and is an upper bound. Suppose $x_1, \dots, x_n \in A$ are linearly dependent. Then $x_1 \in A_{i_1}, \dots, x_n \in A_{i_n}$, so all x_i lie in some A_k as the A_i are a chain. But A_k is linearly independent, which is a contradiction. □

Remark. The only time that linear algebra was used was to show that the maximal element obtained by Zorn's lemma performs the required task; this is usual for proofs in this style.

We can now prove the completeness theorem for propositional logic with no restrictions on the size of the set of primitive propositions.

Theorem. Let $S \subseteq L = L(P)$ be consistent. Then S has a model.

Proof. We will extend S to a consistent set $\overline{S}$ such that for all $t \in L$, either $t \in \overline{S}$ or $\neg t \in \overline{S}$; we then complete the proof by defining a valuation v such that $v(t) = 1$ if $t \in \overline{S}$.

Let $X = \{T \supseteq S \mid T \text{ consistent}\}$ be the poset of consistent extensions of S, ordered by inclusion. We seek a maximal element of X. Then, if $\overline{S}$ is maximal and $t \notin \overline{S}$, then $\overline{S} \cup \{t\} \vdash \bot$ by maximality, so $\overline{S} \vdash \neg t$ by the deduction theorem, giving $\neg t \in \overline{S}$ again by maximality.

Note that $X \neq \varnothing$ as $S \in X$. Given a nonempty chain $(T_i)_{i\in I}$, let $T = \bigcup_{i\in I} T_i$. We have $T \supseteq T_i$ for all i and $T \supseteq S$ as the chain is nonempty, so it suffices to show T is consistent. Indeed, suppose $T \vdash \bot$. Then there exists a subset $\{t_1, \dots, t_n\} \in T$ with $\{t_1, \dots, t_n\} \vdash \bot$ as proofs are finite. Now, $t_1 \in T_{i_1}, \dots, t_n \in T_{i_n}$ so all t_j are elements of T_{i_k} for some k. But T_{i_k} is consistent, so $\{t_1, \dots, t_n\} \nvdash \bot$, giving a contradiction. □

3.3. Well-ordering principle

Theorem. Every set has a well-ordering.

There exist sets with no definable well-ordering, such as $\mathbb{R}$.

Proof. Let S be a set, and let X be the set of pairs (A, R) such that $A \subseteq S$ and R is a well-ordering on A. We define the partial order on X by $(A, R) \leq (A', R')$ if (A', R') extends (A, R), so $R'|_A = A$ and A is an initial segment of A' for R'.

X is nonempty as the empty relation is a well-ordering of the empty set. Given a nonempty chain $(A_i, R_i)_{i\in I}$, there is an upper bound $\left(\bigcup_{i\in I} A_i, \bigcup_{i\in I} R_i\right)$, because the well-orderings are nested. By Zorn's lemma, there exists a maximal element $(A, R) \in X$.

Suppose $x \in S \setminus A$. Then we can construct the well-ordering on $A \cup \{x\}$ by defining $a < x$ for $a \in A$, contradicting maximality of A. Hence $A = S$, so R is a well-ordering on S. □

3.4. Zorn's lemma and the axiom of choice

In the proof of Zorn's lemma, for each $x \in S$ we chose an arbitrary $x' > x$. This requires potentially infinitely many arbitrary choices. Other proofs, such as that the countable union of countable sets is countable, also required infinitely many choices; in this example, we chose arbitrary enumerations of the countable sets $A_1, A_2, \dots$ at once.

Formally, this process of making infinitely many arbitrary choices is known as the *axiom of choice* AC: if we have a family of nonempty sets, one can choose an element from each one. More precisely, for any family of nonempty sets $(A_i)_{i\in I}$, there is a *choice function* $f : I \to \bigcup_{i\in I} A_i$ such that $f(i) \in A_i$ for all i.

Unlike the other axioms of set theory, the function obtained from the axiom of choice is not uniquely defined. For instance, the axiom of union allows for the construction of $A \cup B$ given A and B, which can be fully described; but applying the axiom of choice to the family $\star \mapsto \{1, 2\}$ could give the choice function $\star \mapsto 1$ or $\star \mapsto 2$.

Use of the axiom of choice gives rise to nonconstructive proofs. In modern mathematics it is sometimes considered useful to note when the axiom of choice is being used. However, many proofs that do not even use the axiom of choice are nonconstructive, such as the proof of existence of transcendentals, or Hilbert's basis theorem that every ideal over $\mathbb{Q}[X_1, \dots, X_n]$ is finitely generated.

Although our proof of Zorn's lemma required the axiom of choice, it is not immediately clear that all such proofs require it. However, it can be shown that Zorn's lemma implies the axiom of choice in the presence of the other axioms of ZF set theory. Indeed, if $(A_i)_{i\in I}$ is a family of sets, we can well-order it using the well-ordering principle, and define the choice function by setting $f(i)$ to be the least element of A_i. Hence, Zorn's lemma, the axiom of choice, and the well-ordering principle are equivalent, given ZF.

AC can be proven trivially in ZF for the case $|I| = 1$, because a set being nonempty means precisely that there exists an element inside it. Clearly, AC holds for all finite index sets in ZF by induction on $|I|$. However, ZF does not prove the most general form of AC.

Zorn's lemma is a difficult lemma to prove from first principles because of its reliance on ordinals and Hartogs' lemma; the use of the axiom of choice does not contribute significantly to its difficulty. The construction and properties of the ordinals did not rely on the axiom of choice. The axiom of choice was only used twice in the section on well-orderings: the fact that in a set that is not well-ordered, there is an infinite decreasing sequence; and the fact that ω_1 is not a countable supremum.

4. Predicate logic

4.1. Languages

Recall that a *group* is a set A equipped with functions $m : A^2 \to A$ of arity 2, and $i : A^1 \to A$ of arity 1, and a constant $e \in A$ which can be viewed as a function $A^0 \to A$ of arity 0, such that a set of axioms hold. A *poset* is a set A equipped with a relation $(\leq) \subseteq A^2$ of arity 2, such that a set of axioms hold. Other algebraic structures can be described in the same way.

Let Ω and Π be disjoint sets of functions and relations, and $\alpha : \Omega \cup \Pi \to \mathbb{N}$ be an arity function. *Variables* are symbols of the form x_i for some $i \in \mathbb{N}$. *Terms* are defined inductively by

(i) each variable is a term;

(ii) if $f \in \Omega$ with $\alpha(f) = n$ and terms $t_1, \dots, t_n$, then $f\ t_1\ \dots\ t_n$ is a term.

The *atomic formulae* are defined inductively by

(i) $\bot$ is an atomic formula;

(ii) for terms s, t, $(s = t)$ is an atomic formula;

(iii) if $\varphi \in \Pi$ with $\alpha(\varphi) = n$ and terms $t_1, \dots, t_n$, then $\varphi(t_1, \dots, t_n)$ is an atomic formula.

The *formulae* are defined inductively by

(i) each atomic formula is a formula;

(ii) if p and q are formulae then $(p \Rightarrow q)$ is a formula;

(iii) if p is a formula and x is a variable, then $(\forall x)p$ is a formula.

The *language* $L = L(\Omega, \Pi, \alpha)$ is the set of formulae.

Example. In the language of groups, $\Omega = \{m, i, e\}$ and $\Pi = \varnothing$ with $\alpha(m) = 2, \alpha(i) = 1, \alpha(e) = 0$. $m(x_1, x_2), m(x_1, i(x_2)), e, m(e, e)$ are examples of terms of the language. $e = m(\ell, e), m(x, y) = m(y, x)$ are atomic formulae.

Example. In the language of posets, $\Omega = \varnothing$ and $\Pi = \{\leq\}$ with $\alpha(\leq) = 2$. $x = y, x \leq y$ are atomic formulae. Technically, $x \leq y$ is written $\leq (x, y)$.

Example. In the language of groups, $(\forall x)(m(x, x) = e)$ is a formula. Another formula is $m(x, x) = e \Rightarrow (\exists y)(m(y, y) = x)$.

Remark. A formula is a certain finite string of symbols; it has no intrinsic semantics. We define $\neg p, p \wedge q, p \vee q$ in the usual way. We define $(\exists x)p$ to mean $\neg(\forall x)(\neg p)$.

A term is *closed* if it contains no variables. For example, $e, m(e, i(e))$ are closed in the language of groups, but $m(x, i(x))$ is not closed.

An occurrence of a variable x in a formula p is *bound* if it is inside the brackets of a $(\forall x)$ quantifier. Otherwise, we say the occurrence is *free*. In the formula $(\forall x)(m(x, x) = e)$, each occurrence of x is bound. In $m(x, x) = e \Rightarrow (\exists y)(m(y, y) = x)$, the occurrences of x are free and the occurrences of y are bound. In the formula $m(x, x) = e \Rightarrow (\forall x)(\forall y)(m(x, y) = m(y, x))$, the occurrences of x on the left hand side are free, and the occurrences of x on the right hand side are bound.

A *sentence* is a formula with no free variables. For instance, $(\forall x)(m(x, x) = e)$ is a sentence, and $(\forall x)(m(x, x) \Rightarrow (\exists y)(m(y, y) = x))$ is a sentence. In the language of posets, $(\forall x)(\exists y)(x \geq y \wedge \neg(x = y))$ is a sentence.

For a formula p, term t, and variable x, the *substitution* $p[t/x]$ is obtained from p by replacing every free occurrence of x with t. For example,

$$p = (\exists y)(m(y, y) = x); \quad p[e/x] = (\exists y)(m(y, y) = e)$$

4.2. Semantic implication

Definition. Let $L = L(\Omega, \Pi, \alpha)$ be a language. An L*-structure* is

- a nonempty set A;
- for each $f \in \Omega$, a function $f_A : A^n \to A$ where $n = \alpha(f)$;
- for each $\varphi \in \Pi$, a subset $\varphi_A \subseteq A^n$ where $n = \alpha(\varphi)$.

Remark. We will see later why the restriction that A is nonempty is given here.

Example. In the language of groups, an L-structure is a nonempty set A with functions $m_A : A^2 \to A, i_A : A \to A, e_A \in A$. Such a structure may not be a group, as we have not placed any axioms on A.

Example. In the language of posets, an L-structure is a nonempty set A with a relation $(\leq_A) \subseteq A^2$.

We define the *interpretation* $p_A \in \{0, 1\}$ of a sentence p in an L-structure A as follows.

- The interpretation t_A of a closed term t in an L-structure A is defined inductively as $(f\ t_1 \dots t_n)_A = f_A(t_{1A}, \dots, t_{nA})$ for $f \in \Omega, \alpha(f) = n$, where $t_1, \dots, t_n$ are closed.
- The interpretation of an atomic sentence is defined inductively.
  - $\bot_A = 0$.
  - $(s = t)_A$ is 1 if $s_A = t_A$ and 0 if $s_A \neq t_A$.
  - $(\varphi(t_1, \dots, t_n))_A$ is 1 if $(t_{1A}, \dots, t_{nA}) \in \varphi_A$ and 0 otherwise, for $\varphi \in \Pi, \alpha(\varphi) = n$, where $t_1, \dots, t_n$ are closed.
- We now inductively define the interpretation of sentences, which is technically induction by length over all languages at once.
  - $(p \Rightarrow q)_A$ is 0 if $p_A = 1$ and $q_A = 0$, and 1 otherwise.

- $((\forall x)p)_A$ is 1 if $p[\overline{a}/x]$ is 1 for all $a \in A$ and 0 otherwise, where we add a constant symbol $\overline{a}$ to L for a fixed $a \in A$ to form the language L', and we make A into an L'-structure by defining $\overline{a}_A = a$.

Remark. For a formula p with free variables, we can define p_A to be the subset of A^k where k is the number of free variables, defined such that $x \in p_A$ if and only if the substitution of x in p is evaluated to 1.

Definition. If $p_A = 1$, we say p *holds* in A, or p is *true* in A, or A is a *model* of p. A *theory* is a set of sentences, known as its *axioms*. We say that A is a *model* of a theory T if $p_A = 1$ for all $p \in T$. For a theory T and a sentence p, we say that $T \vDash p$, read T *entails* or *semantically implies* p, if every model of T is a model of p.

Example. Let L be the language of groups, and let

$$\begin{aligned} T = \{&(\forall x)(\forall y)(\forall z)(m(x, m(y,z)) = m(m(x,y),z)), \\ &(\forall x)(m(x,e) = x \wedge m(e,x) = x), \\ &(\forall x)(m(x,i(x)) = e \wedge m(i(x),x) = e)\} \end{aligned}$$

Then, an L-structure is a model of T if and only if it is a group. Note that this statement has two assertions; every L-structure that is a model of T is a group, and that every group can be turned into an L-structure that models T. We say that T *axiomatises* the theory of groups or the class of groups.

Example. Let L be the language of posets, and T be the poset axioms. Then T axiomatises the class of posets.

Example. Let L be the language of fields, so $\Omega = \{0, 1, +, \cdot, -\}$ with $\alpha(0) = \alpha(1) = 0, \alpha(+) = \alpha(\cdot) = 2, \alpha(-) = 1$. T is the usual field axioms, including the statement $(\forall x)(\neg(x = 0) \Rightarrow (\exists y)(x \cdot y = 1))$. Then T entails the statement that inverses are unique: $(\forall x)(\neg(x = 0) \Rightarrow (\forall y)(\forall z)(y \cdot x = 1 \wedge z \cdot x = 1 \Rightarrow y = z))$.

Example. Let L be the language of graphs, defined by $\Omega = \varnothing$ and $\Pi = \{a\}$ where $\alpha(a) = 2$ is the adjacency relation. Define $T = \{(\forall x)(\neg a(x,x)), (\forall x)(\forall y)(a(x,y) \Rightarrow a(y,x))\}$. Then T axiomatises the class of graphs.

4.3. Syntactic implication

We need to define (logical) axioms and deduction rules in order to construct proofs.

(i) $p \Rightarrow (q \Rightarrow p)$ for formulae p, q.

(ii) $(p \Rightarrow (q \Rightarrow r)) \Rightarrow ((p \Rightarrow q) \Rightarrow (p \Rightarrow r))$ for formulae p, q, r.

(iii) $\neg\neg p \Rightarrow p$ for each formula p.

(iv) $(\forall x)(x = x)$ for any variable x.

(v) $(\forall x)(\forall y)(x = y \Rightarrow (p \Rightarrow p[y/x]))$ for any variables x, y where y is not bound in the formula p.

(vi) $((\forall x)p) \Rightarrow p[t/x]$ for any variable x, formula p, and term t that has no free variable that occurs bound in p.

(vii) $(\forall x)(p \Rightarrow q) \Rightarrow (p \Rightarrow (\forall x)q)$ for any formulae p, q and variable x that does not appear free in p.

Note that all of these axioms are tautologies; they hold in every structure. We define the following deduction rules.

(i) (modus ponens) From p and $p \Rightarrow q$, we can deduce q.

(ii) (generalisation) From p, we can deduce $(\forall x)p$ provided that x does not occur free in any premise used to deduce p.

For $S \subseteq L$ and $t \in L$, we say that $S \vdash p$, read S *proves* p, if there exists a *proof* of p from S, which is a finite sequence of formulae ending with p such that each formula is a logical axiom, a hypothesis in S, or obtained from earlier lines by one of the deduction rules.

Remark. Suppose we allow the empty structure for a language with no constants. Then, $\bot$ is false in A, and the statement $(\forall x)\bot$ is true in A. Therefore, $((\forall x)\bot) \Rightarrow \bot$ is false by modus ponens. But this is an instance of axiom (vi), showing that it would not be a tautology.

Example. We show $\{x = y, x = z\} \vdash y = z$ where x, y, z are different variables.

1. $(\forall x)(\forall y)(x = y \Rightarrow (x = z \Rightarrow y = z))$ (axiom 5)
2. $((\forall x)(\forall y)(x = y \Rightarrow (x = z \Rightarrow y = z))) \Rightarrow (\forall y)(x = y \Rightarrow (x = z \Rightarrow y = z))$ (axiom 6)
3. $(\forall y)(x = y \Rightarrow (x = z \Rightarrow y = z))$ (modus ponens on lines 1, 2)
4. $((\forall y)(x = y \Rightarrow (x = z \Rightarrow y = z))) \Rightarrow (x = y \Rightarrow (x = z \Rightarrow y = z))$ (axiom 6)
5. $x = y \Rightarrow (x = z \Rightarrow y = z)$ (modus ponens on lines 3, 4)
6. $x = y$ (hypothesis)
7. $x = z \Rightarrow y = z$ (modus ponens on lines 5, 6)
8. $x = z$ (hypothesis)
9. $y = z$ (modus ponens on lines 7, 8)

4.4. Deduction theorem

Proposition. Let $S \subseteq L$, and $p, q \in L$. Then $S \vdash (p \Rightarrow q)$ if and only if $S \cup \{p\} \vdash q$.

Proof. As before, given a proof of $p \Rightarrow q$ from S, one can establish a proof of q from $S \cup \{p\} \vdash q$ by writing p and applying modus ponens to the original proof.

Conversely, suppose we have a proof $S \cup \{p\} \vdash q$. We convert each line t_i into $p \Rightarrow t_i$ as in the proof in propositional logic. The only new case is generalisation. Suppose we have the line r and then the line $(\forall x)r$ obtained by generalisation, and we have a proof $S \vdash p \Rightarrow r$ by induction. In the proof $S \cup \{p\} \vdash r$, no hypothesis has a free occurrence of x. Therefore, in the proof $S \vdash p \Rightarrow r$, the same holds. Thus, $S \vdash (\forall x)(p \Rightarrow r)$ by generalisation.

Suppose x is not free in p. Then, $S \vdash p \Rightarrow (\forall x)r$ by axiom 7 and modus ponens.

Now, suppose x occurs free in p. In this case, the proof $S \cup \{p\} \vdash r$ cannot have used the hypothesis p. Hence, $S \vdash r$, and so $S \vdash (\forall x)r$ by generalisation. This gives $S \vdash p \Rightarrow (\forall x)r$ by axiom 1. □

4.5. Soundness

This section is non-examinable.

Proposition. Let S be a set of sentences in L, and p a sentence in L. Then $S \vdash t$ implies $S \vDash t$.

Proof. We have a proof $t_1, \dots, t_n$ of p from S. We show that if A is a model of S, A is also a model of t_i for each i (interpreting free variables as quantified); this can be shown by induction. Hence, $S \vDash p$. □

4.6. Adequacy

This section is non-examinable.

We want to show that $S \vDash p$ implies $S \vdash p$. Equivalently, $S \cup \{\neg p\} \vDash \bot$ implies $S \cup \{\neg p\} \vdash \bot$. In other words, if $S \cup \{\neg p\}$ is consistent, it has a model.

Theorem (model existence lemma)**.** Every consistent theory has a model.

We will need a number of key ideas in order to prove this.

(i) We will construct our model out of the language itself using the closed terms of L. For instance, if L is the language of fields and S is the usual field axioms, we take the closed terms and combine them with $+$ and $\cdot$ in the obvious way.

(ii) However, we can prove $S \vdash 1 + 0 = 1$, but $1 + 0$ and 1 are distinct as strings. We will therefore take the quotient of this set by the equivalence relation defined by $s \sim t$ if $S \vdash s = t$. If this set is A, we define $[s] +_A [t] = [s + t]$, and this is a well-defined operation.

(iii) Suppose S is the set of field axioms with the statement that $1 + 1 = 0 \vee 1 + 1 + 1 = 0$. In this theory, $S \nvdash 1 + 1 = 0$ and $S \nvdash 1 + 1 + 1 = 0$. Therefore, $[1 + 1] \neq [0]$ and $[1 + 1 + 1] \neq [0]$, so our structure A is not of characteristic 2 or 3. We can overcome this by first extending S to a maximal consistent theory.

(iv) Suppose S is the set of field axioms with the statement that $(\exists x)(x \cdot x = 1 + 1)$. There is no closed term t with the property that $[t \cdot t] = [1 + 1]$. The problem is that S lacks *witnesses* to existential quantifiers. For each statement of the form $(\exists x)p \in S$, we add a new constant c to the language and add to S the sentence $p[c/x]$. This still forms a consistent set.

(v) The resulting set may no longer be maximal, as we have extended our language with new constants. We must then return to step (iii) then step (iv); it is not clear if this process ever terminates.

Proof. Let S be a consistent set in a language $L = L(\Omega, \Pi)$. Extend S to a maximal consistent set S_1, using Zorn's lemma. Then, for each sentence $p \in L$, either $p \in S_1$ or $\neg p \in S_1$. Such a theory is called *complete*; each sentence or its negation is proven. Now, we add witnesses to S_1: for each sentence of the form $(\exists x)p \in S_1$, we add a new constant symbol c to the language, and also add the sentence $p[c/x]$. We then obtain a new theory T_1 in the language $L_1 = L(\Omega \cup C_1\Pi)$ that has witnesses for every existential in S_1. One can check easily that T_1 is consistent.

We then extend T_1 to a maximal consistent theory S_2 in L_1, and add witnesses to produce T_2 in the language $L_2 = L(\Omega \cup C_1 \cup C_2, \Pi)$. Continue inductively, and let $\overline{S} = \bigcup_{n\in\mathbb{N}} S_n$ in the language $\overline{L} = L\left(\Omega \cup \bigcup_{n\in\mathbb{N}} C_n, \Pi\right)$.

We claim that $\overline{S}$ is consistent, complete, and has witnesses for every existential in $\overline{S}$. Clearly $\overline{S}$ is consistent: if $\overline{S} \vdash \bot$ then $S_n \vdash \bot$ for some n as proofs are finite, contradicting consistency of S_n. For completeness, if p is a sentence in $\overline{L}$, p must lie in L_n for some n as it is a finite string of symbols. But S_{n+1} is complete in L_n, so $S_{n+1} \vdash p$ or $S_{n+1} \vdash \neg p$, so certainly $\overline{S} \vdash p$ or $\overline{S} \vdash \neg p$. If $(\exists x)p \in \overline{S}$, then $(\exists x)p \in S_n$ for some n, so T_n provides a witness.

On the closed terms of $\overline{L}$, we define the relation $s \sim t$ if $\overline{S} \vdash s = t$. This is clearly an equivalence relation, so we can define A to be the set of equivalence classes of $\overline{L}$ under $\sim$. This is an $\overline{L}$-structure by defining

- $f_A([t_1], \dots, [t_n]) = [f\ t_1 \dots t_n]$ for each $f \in \Omega \cup \bigcup_{n\in\mathbb{N}} C_n, \alpha(f) = n, t_i$ closed terms;
- $\varphi_A = \left\{([t_1], \dots, [t_n]) \in A^n \mid \overline{S} \vdash \varphi(t_1, \dots, t_n)\right\}$ for each $\varphi \in \Pi, \alpha(\varphi) = n, t_i$ closed terms.

We claim that for a sentence $p \in \overline{L}$, we have $p_A = 1$ if and only if $\overline{S} \vdash p$. Then the proof is complete, as $S \subseteq \overline{S}$ so $p_A = 1$ for every $p \in S$, so A is a model of S.

We prove this by induction on the length of sentences. First, suppose p is atomic. $\bot_A = 0$, as $\overline{S} \nvdash \bot$. For closed terms s, t, $\overline{S} \vdash s = t$ if and only if $[s] = [t]$ by definition of $\sim$. This holds if and only if $s_A = t_A$ by definition of the operations in A. This is precisely the statement that $s = t$ holds in A. The same holds for relations.

Now consider $p \Rightarrow q$. $\overline{S} \vdash p \Rightarrow q$ if and only if $\overline{S} \vdash \neg p$ or $\overline{S} \vdash q$ as $\overline{S}$ is complete and consistent; if $\overline{S} \nvdash \neg p$ and $\overline{S} \nvdash q$, then $\overline{S} \vdash p$ and $\overline{S} \vdash \neg p$. By induction on the length of the formula, this holds if and only if $p_A = 0$ or $q_A = 1$. This is the definition of the interpretation of $p \Rightarrow q$ in A.

Finally, consider the existential $(\exists x)p$. $\overline{S} \vdash (\exists x)p$ if and only if there is a closed term t such that $\overline{S} \vdash p[t/x]$, as $\overline{S}$ has witnesses to every existential. By induction (for example on the amount of quantifiers in a formula), this holds if and only if $p[t/x]_A = 1$ for some closed term t. This is true exactly when $(\exists x)p$ holds in A, as A is precisely the set of equivalence classes of closed terms. □

Corollary (adequacy). Let $S \subseteq L$ be a theory and $t \in L$ be a sentence. Then $S \vDash t$ implies $S \vdash t$.

4.7. Completeness

Theorem (Gödel's completeness theorem for first order logic). Let $S \subseteq L$ be a theory and $t \in L$ be a sentence. Then $S \vDash t$ if and only if $S \vdash t$.

Proof. Follows from soundness and adequacy. □

Note that *first order* refers to the fact that variables quantify over elements, rather than sets of elements.

Remark. If L is countable, or equivalently Ω and Π are countable, Zorn's lemma is not needed in the above proof.

Theorem (compactness theorem). Let $S \subseteq L$ be a theory. Then if every finite subset $S' \subseteq S$ has a model, S has a model.

Proof. Trivial after applying completeness as proofs are finite. □

There is no decidability theorem for first order logic, as $S \vDash p$ can only be verified by checking its valuation in every L-structure.

Corollary. The class of finite groups is not axiomatisable in the language of groups: there is no theory S such that a group is finite if and only if each $p \in S$ holds in the group.

Proof. Suppose S is a set of sentences that axiomatises the theory of finite groups. Consider S together with the sentences $(\exists x_1)(\exists x_2)(x_1 \neq x_2)$, $(\exists x_1)(\exists x_2)(\exists x_3)(x_1 \neq x_2 \wedge x_1 \neq x_3 \wedge x_2 \neq x_3)$ and so on, which collectively assert that the group has at least k elements for every k. Each finite subset $S' \subseteq S$ has a model, such as a cyclic group of sufficiently large order. So by compactness, there is a model of S, which is a finite group with at least k elements for every k, giving a contradiction. □

Corollary. Let S be a theory with arbitrarily large finite models. Then S has an infinite model.

Proof. Add sentences and apply compactness as in the previous corollary. □

Finiteness is not a first-order property.

Theorem (upward Löwenheim–Skolem theorem)**.** Let S be a theory with an infinite model. Then S has an uncountable model.

Proof. Add constants $\{c_i \mid i \in I\}$ to the language, where I is an uncountable set. Add sentences $c_i \neq c_j$ to the theory for all $i \neq j$ to obtain a theory S'. Any finite set of sentences in S' has a model: indeed, the infinite model of S suffices. By compactness, S' has a model. □

Remark. Similarly, we can prove the existence of models of S that do not inject into X for any fixed set X. Adding $\gamma(X)$ constants or $\mathcal{P}(X)$ constants both suffice.

Example. There is an uncountable field, as there is an infinite field $\mathbb{Q}$. There is also a field that does not inject into X for any fixed set X.

Theorem (downard Löwenheim–Skolem theorem)**.** Let S be a theory in a countable language L, or equivalently, Ω and Π are countable. Then if S has a model, it has a countable model.

Proof. S is consistent, so the model constructed in the proof of the model existence lemma is countable. □

4.8. Peano arithmetic

Consider the language L given by $\Omega = \{0, s, +, \cdot\}$ with $\alpha(0) = 0, \alpha(s) = 1, \alpha(+) = \alpha(\cdot) = 2$, and $\Pi = \varnothing$. It has axioms

(i) $(\forall x)(s(x) \neq 0)$;

(ii) $(\forall x)(\forall y)(s(x) = s(y) \Rightarrow x = y)$;

(iii) $(\forall y_1) \dots (\forall y_n)[p[0/x] \wedge (\forall x)(p \Rightarrow p[s(x)/x]) \Rightarrow (\forall x)p]$ for each formula p with free variables $x, y_1, \dots, y_n$;

(iv) $(\forall x)(x + 0 = x)$;

(v) $(\forall x)(\forall y)(x + s(y) = s(x + y))$;

(vi) $(\forall x)(x \cdot 0 = 0)$;

(vii) $(\forall x)(\forall y)(x \cdot s(y) = x \cdot y + x)$.

These axioms are sometimes called Peano arithmetic, PA, or formal number theory. The y_i in (iii) are called *parameters*. Without the parameters, we would not be able to perform induction on sets such as $\{x \mid x \geq y\}$ if y is a variable.

Note that PA clearly has an infinite model, namely $\mathbb{N}$. So by the upward Löwenheim–Skolem theorem, it has an uncountable model, which in particular is not isomorphic to $\mathbb{N}$. This is because (iii) is not 'true' induction, stating that all subsets of $\mathbb{N}$ either have a least element not in it, or it is $\mathbb{N}$ itself. Axiom (iii) applies only to countably many formulae p, and therefore only asserts that induction holds for countably many subsets of $\mathbb{N}$.

Definition. A set $S \subseteq \mathbb{N}$ is *definable* in the language of PA if there is a formula p with a free variable x such that for each $m \in \mathbb{N}$, $m \in S$ if and only if $p[m/x]$ holds in $\mathbb{N}$.

Only countably many formulae exist, so only countably many sets are definable.

Example. The set of squares is definable, as it can be defined by the formula $(\exists y)(y \cdot y = x)$. The set of primes is also definable by $x \neq 0 \wedge x \neq 1 \wedge (\forall y)(y \mid x \Rightarrow y = 1 \wedge y = x)$, where $y \mid x$ is defined to mean $(\exists z)(z \cdot y = x)$. The set of powers of 2 can be defined by $(\forall y)(y \text{ is prime} \wedge y \mid x \Rightarrow y = 2)$. The set of powers of 4 and the set of powers of 6 are also definable.

Theorem (Gödel's incompleteness theorem)**.** PA is not complete.

This theorem shows that there is a sentence p such that PA $\nvdash p$ and PA $\nvdash \neg p$. However, one of $p, \neg p$ must hold in $\mathbb{N}$, so there is a sentence p that is true in $\mathbb{N}$ that PA does not prove. This does not contradict the completeness theorem, which is that if p is true in *every* model in PA then PA $\vdash p$.

5. Set theory

5.1. Axioms of ZF

In this section, we will attempt to understand the structure of the universe of sets. In order to do this, we will treat set theory as a first-order theory like any other, and can therefore study it with our usual tools. In particular, we will study a particular theory called *Zermelo–Fraenkel set theory*, denoted ZF. The language has $\Omega = \varnothing, \Pi = \{\in\}, \alpha(\in) = 2$. A ‘universe of sets’ is simply a model $(V, \in_V) = (V, \in)$ for the axioms of ZF. We can view this section as a worked example of the concepts of predicate logic, but every model of ZF will contain a copy of (most of) mathematics, so they will be very complicated.

We now define the axioms of ZF set theory.

(i) *Axiom of extension.*

$$(\forall x)(\forall y)((\forall z)(z \in x \Leftrightarrow z \in y) \Rightarrow x = y)$$

Note that the converse follows from the definition of equality. This implies that sets have no duplicate elements, and have no ordering.

(ii) *Axiom of separation* or *comprehension.* For a set x and a property p, we can form the set of $z \in x$ such that $p(z)$ holds.

$$(\forall t_1) \dots (\forall t_n)(\forall x)(\exists y)(\forall z)(z \in y \Leftrightarrow z \in x \wedge p)$$

where the t_i are the parameters, and p is a formula with free variables $t_1, \dots, t_n, z$. Note that we need the parameters as we may wish to form the set $\{z \in x \mid z \in t\}$ for some variable t. We write $\{z \in x \mid p(z)\}$ for the set guaranteed by this axiom; this is an abbreviation and does not change the language.

(iii) *Empty-set axiom.*

$$(\exists x)(\forall y)(\neg y \in x)$$

This empty set is unique by extensionality. We write $\varnothing$ for the set guaranteed by this axiom. For instance, $p(\varnothing)$ is the sentence $(\exists x)((\forall y)(\neg y \in x) \wedge p(x))$.

(iv) *Pair-set axiom.*

$$(\forall x)(\forall y)(\exists z)(\forall t)(t \in z \Leftrightarrow t = x \vee t = y)$$

We write $\{x, y\}$ for this set z, which is unique by extensionality. Some basic set-theoretic principles can now be defined.

- We write $\{x\} = \{x, x\}$ for the singleton set containing x.
- We can now define the ordered pair $(x, y) = \{\{x\}, \{x, y\}\}$; from the axioms so far we can prove that $(x, y) = (z, t)$ if and only if $x = z$ and $y = t$.
- We say that x is an ordered pair if $(\exists y)(\exists z)(x = (y, z))$, and f is a function if

$$(\forall x)(x \in f \Rightarrow x \text{ is an ordered pair})$$

and

$$(\forall x)(\forall y)(\forall z)((x,y) \in f \wedge (x,z) \in f \Rightarrow y = z)$$

- We call a set x the domain of f, written $x = \operatorname{dom} f$, if f is a function and

$$(\forall y)(y \in x \Leftrightarrow (\exists z)((y,z) \in f))$$

- The notation $f : x \to y$ means that f is a function, $x = \operatorname{dom} f$, and

$$(\forall z)(\forall t)((z,t) \in f \Rightarrow t \in y)$$

(v) *Union axiom.* For each family of sets x, we can form its union $\bigcup_{t \in x} t$.

$$(\forall x)(\exists y)(\forall z)(z \in y \Leftrightarrow (\exists t)(z \in t \wedge t \in x))$$

The set guaranteed by this axiom can be written $\bigcup x$, and we can write $x \cup y$ for $\bigcup \{x, y\}$. We need no intersection axiom, as such intersections already exist by the axiom of separation. This cannot be used to create empty intersections, as the axiom of separation can only create subsets of a set that already exists.

(vi) *Power-set axiom.*

$$(\forall x)(\exists y)(\forall z)(z \in y \Leftrightarrow z \subseteq x)$$

where $z \subseteq x$ means $(\forall t)(t \in z \Rightarrow t \in x)$. We write $\mathcal{P}(x)$ for the power set of x. We can form the Cartesian product $x \times y$ as a suitable subset of $\mathcal{P}(\mathcal{P}(x \cup y))$, as if $z \in x, t \in y$, we have $(z,t) = \{\{z\},\{z,t\}\} \in \mathcal{P}(\mathcal{P}(x \cup y))$. The set of all functions $x \to y$ can be defined as a subset of $\mathcal{P}(x \times y)$.

(vii) *Axiom of infinity.* Using our currently defined axioms, any model V must be infinite. For example, writing x^+ for the *successor* of x defined as $x \cup \{x\}$, the sets $\varnothing, \varnothing^+, \varnothing^{++}, \dots$ are distinct.

$$\varnothing^+ = \{\varnothing\}; \quad \varnothing^{++} = \{\varnothing, \{\varnothing\}\}; \quad \varnothing^{+++} = \{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}\}; \quad \dots$$

We write $0 = \varnothing, 1 = \varnothing^+, 2 = \varnothing^{++}, \dots$ for the successors created in this way. For instance, $3 = \{0, 1, 2\}$. V may not have an infinite element, even though V itself is infinite, because no $x \in V$ has all $y \in V$ as elements: V does not think of itself as a set, because Russell's paradox follows from the axioms defined so far.

We say that x is a successor set if $\varnothing \in x$ and $(\forall y)(y \in x \Rightarrow y^+ \in x)$. Note that this is a finite-length formula that characterises an infinite set. The axiom of infinity is that there exists a successor set.

$$(\exists x)(\varnothing \in x \wedge (\forall y)(y \in x \Rightarrow y^+ \in x))$$

Note that this set is not uniquely defined, but any intersection of successor sets is a successor set. We can therefore take the intersection of all successor sets by the axiom of separation, giving a least successor set denoted ω. Thus, $(\forall x)(x \in \omega \Leftrightarrow (\forall y)(y \text{ is a successor set} \Rightarrow x \in y))$. For example, we can prove that $3 \in \omega$.

In particular, if x is a successor set and a subset of ω, then $x = \omega$. Hence, $(\forall x)(x \subseteq \omega \wedge \varnothing \in x \wedge (\forall y)(y \in x \Rightarrow y^+ \in x) \Rightarrow x = \omega)$. This is 'proper' induction over all subsets of ω, unlike the weaker first-order induction defined in the Peano axioms. It is easy to check that $(\forall x)(x \in \omega \Rightarrow x^+ \neq \varnothing)$ and $(\forall x)(\forall y)(x \in \omega \wedge y \in \omega \wedge x^+ = y^+ \Rightarrow x = y)$, so ω satisfies (in V) the usual axioms for the natural numbers. We can now define 'x is finite' to mean $(\exists y)(y \in \omega \wedge x$ bijects with $y)$, and define 'x is countable' to mean that x is finite or bijects with ω.

(viii) *Axiom of foundation* or *regularity*. We require that sets are built out of simpler sets. For example, we want to disallow a set from being a member of itself, and similarly forbid $x \in y$ and $y \in x$. In general, we want to forbid sets x_i such that $x_{i+1} \in x_i$ for each $i \in \mathbb{N}$.

Note that if $x \in x$, $\{x\}$ has no $\in$-minimal element. If $x \in y, y \in x$, $\{x, y\}$ has no $\in$-minimal element. In the last example, $\{x_0, x_1, \dots\}$ has no $\in$-minimal element. We now define the axiom of foundation: every nonempty set has an $\in$-minimal element.

$$(\forall x)(x \neq \varnothing \Rightarrow (\exists y)(y \in x \wedge (\forall z)(z \in x \Rightarrow z \notin y)))$$

Any model of ZF without this axiom has a submodel of all of ZF.

(ix) *Axiom of replacement*. Often, we are given an index set I and construct a set A_i for each $i \in I$, then take the collection $\{A_i \mid i \in I\}$. In order to write this down, the mapping $i \mapsto A_i$ must be a function, or equivalently, there must be a set $\{(i, A_i) \mid i \in I\}$. This is not clear from the other axioms. We would like to say that the image of a set under something that looks like a function (since we do not yet have such a set-theoretic function) is a set.

Let $(V, \in)$ be an L-structure. A *class* is a set $C \subseteq V$ such that for some formula p with free variables x and some parameters, we have $x \in C$ if and only if p holds in V. C is a set outside of our model; it may not correspond to a set $x \in V$ inside the model. For instance, V is a class, taking p to be $x = x$. There is a class of infinite sets, taking p to be 'x is not finite'. For any $t \in V$, the collection of x with $t \in x$ is a class; here, t is a parameter to the class. Every set $y \in V$ is a class by setting p to be $x \in y$. A *proper class* is a class that does not correspond to a set $x \in V$: $\neg(\exists y)(\forall x)(x \in y \Leftrightarrow p)$. When writing about classes inside ZF, we instead write about their defining formulae, as classes have no direct representation in the language.

Similarly, a *function-class* is a set $F \subseteq V$ of ordered pairs from V such that for some formula p with free variables x, y and parameters, we have (x, y) belongs to F if and only if p, and if $(x, y), (x, z)$ belong to F, $y = z$. This is intuitively a function whose domain may not be a set. For example, the mapping $x \mapsto \{x\}$ is a function-class, taking p to be $y = \{x\}$. This is not a function, for example, every f has a domain which is a set in V, and this function has domain V which is not a set.

We can now define the axiom of replacement: the image of a set under a function-class is a set.

$$(\forall t_1) \dots (\forall t_n)[(\forall x)(\forall y)(\forall z)(p \wedge p[z/y] \Rightarrow y = z) \Rightarrow$$
$$(\forall x)(\exists y)(\forall z)(z \in y \Leftrightarrow (\exists t)(t \in x \wedge p[t/x, z/y]))]$$

For example, for any set x, we can form the set $\{\{t\} \mid t \in x\}$, which is the image of x under the function class $t \mapsto \{t\}$. This set could alternatively have been formed using the power-set and separation axioms; we will later present some examples of sets built with this axiom that cannot be constructed from the other axioms.

This completes the description of the axioms of ZF. We write ZFC for ZF + AC, where AC is the axiom

$$(\forall f)[f \text{ is a function} \wedge (\forall x)(x \in \operatorname{dom} f \Rightarrow f(x) \neq \varnothing) \Rightarrow$$
$$(\exists g)(g \text{ is a function} \wedge (\operatorname{dom} g = \operatorname{dom} f) \wedge (\forall x)(x \in \operatorname{dom} f \Rightarrow g(x) \in f(x)))]$$

5.2. Transitive sets

Definition. x is *transitive* if each member of a member of x is a member of x.

$$(\forall y)((\exists z)(y \in z \wedge z \in x) \Rightarrow y \in x)$$

Equivalently, $\bigcup x \subseteq x$.

Example. $\varnothing$ is a transitive set. $\{\varnothing\}$ is also transitive, and $\{\varnothing, \{\varnothing\}\}$ is transitive. In general, elements of ω are transitive. This can be proven by ω-induction (inside a model): $\varnothing$ is transitive, and if y is transitive, $y^+ = y \cup \{y\}$ is clearly transitive.

Lemma. Every set is contained in a transitive set.

Here, we define 'x contains y' to mean $y \subseteq x$, not $y \in x$.

Remark. This proof takes place inside an arbitrary model of ZF. Technically, the statement of the lemma is 'let $(V, \in)$ be a model of ZF, then for all sets $x \in V$, x is contained in a transitive set $y \in V$'. By completeness, this will show that there is a proof of this fact from the axioms of ZF.

Note also that once this lemma is proven, any x is contained in a least transitive set by taking intersections, called its *transitive closure*, written $TC(x)$. This holds as any intersection of transitive sets is transitive.

Proof. We want to form $x \cup (\bigcup x) \cup (\bigcup \bigcup x) \cup \dots$; if this is a set, it is clearly transitive and contains x. We can show that this is a set by the union axiom applied to the set $\{x, \bigcup x, \bigcup \bigcup x, \dots\}$. This is a set by applying the axiom of replacement, it is an image of ω under the function-class $0 \mapsto x, 1 \mapsto \bigcup x, 2 \mapsto \bigcup \bigcup x$ and so on. We want to define the function-class $p(z, w)$ to be $(z = 0 \wedge w = x) \vee ((\exists t)(\exists u) z = t^+ \wedge w = \bigcup u \wedge p(t, u))$, but this is not a first-order formula.

Define that f is an *attempt* to mean that

$$(f \text{ is a function}) \wedge (\operatorname{dom} f \in \omega) \wedge (\operatorname{dom} f \neq \varnothing) \wedge (f(0) = x) \wedge$$
$$(\forall n)\big(n \in \omega \wedge n \in \operatorname{dom} f \wedge n \neq 0 \Rightarrow f(n) = \bigcup f(n-1)\big)$$

Then,

$$(\forall n)(n \in \omega \Rightarrow (\exists f)(f \text{ is an attempt} \wedge n \in \operatorname{dom} f))$$

can be proven by ω-induction. We can similarly prove

$$(\forall n)(n \in \omega \Rightarrow (\forall f)(\forall g)(f, g \text{ are attempts} \wedge n \in \operatorname{dom} f \cap \operatorname{dom} g \Rightarrow f(n) = g(n)))$$

by ω-induction. We now define the function-class $p = p(z, w)$ to be

$$(\exists f)(f \text{ is an attempt} \wedge z \in \operatorname{dom} f \wedge f(z) = w)$$

□

Intuitively, we needed to use the axiom of replacement because we started with a set x and needed to go 'far away' from it, forming $\bigcup^n x$ for all x. We could not have used the other axioms such as the power-set axiom, as the $\bigcup^n x$ are not contained in an obvious larger set.

Transitive closures allow us to pass from the large universe of sets, which is not a set itself, into a smaller world which is a set closed under $\in$ that contains the relevant sets in question.

5.3. $\in$-induction

We want the axiom of foundation to capture the idea that sets are built out of simpler sets.

Theorem (principle of $\in$-induction)**.** For each formula p with free variables $t_1, \dots, t_n, x$,

$$(\forall t_1) \dots (\forall t_n)[(\forall x)((\forall y)(y \in x \Rightarrow p(y)) \Rightarrow p(x)) \Rightarrow (\forall x)p(x)]$$

Proof. Given $t_1, \dots, t_n$ and the statement $(\forall x)((\forall y)(y \in x \Rightarrow p(y)) \Rightarrow p(x))$, we want to show $(\forall x)p(x)$. Suppose this is not the case, so there exists x such that $\neg p(x)$. We want to look at the set $\{t \mid \neg p(t)\}$ and take an $\in$-minimal element, but this is not necessarily a set, for instance if $p(x)$ is the assertion $x \neq x$.

Let $u = \{t \in TC(\{x\}) \mid \neg p(t)\}$; this is clearly a set in the model, and $u \neq \varnothing$ as $x \in u$. Let t be an $\in$-minimal element of u, guaranteed by the axiom of foundation. Then $\neg p(t)$ as $t \in u$, but $p(z)$ for all $z \in t$ by minimality of t, noting that $z \in t$ implies $z \in TC(\{x\})$. This gives a contradiction. □

The name of this theorem should be read 'epsilon-induction', even though the membership relation is denoted $\in$ and not ϵ or ε.

The principle of $\in$-induction is equivalent to the axiom of foundation in the presence of the other axioms of ZF. We say that x is *regular* if $(\forall y)(x \in y \Rightarrow y$ has a minimal element). The axiom of foundation is equivalent to the assertion that every set is regular. Given $\in$-induction, we can prove every set is regular. Suppose $(\forall y \in x)(y$ is regular); we need to show x is regular. For a set z with $x \in z$, if x is minimal in z, x is clearly regular as required. If x is not minimal in z, there exists $y \in x$ such that $y \in z$. So z has a minimal element as y is regular. Hence x is regular.

5.4. $\in$-recursion

We want to be able to define $f(x)$ given $f(y)$ for all $y \in x$.

Theorem ($\in$-recursion theorem). Let G be a function-class, so $(x, y) \in G$ if and only if $p(x, y)$ for some formula p. Suppose that G is defined for all sets. Then there is a function-class F defined for all sets by a formula $q(x, y)$ such that

$$(\forall x)\left(F(x) = G\left(F\Big|_x\right)\right)$$

Moreover, this F is unique.

Note that $F|_x = \{(y, F(y)) \mid y \in x\}$ is a set by the axiom of replacement.

Proof. Define that f is an *attempt* if

$$f \text{ is a function} \wedge \operatorname{dom} f \text{ is transitive} \wedge (\forall x)\left(x \in \operatorname{dom} f \Rightarrow f(x) = G\left(f\Big|_x\right)\right)$$

Note that $f|_x$ is defined as $\operatorname{dom} f$ is transitive. Then,

$$(\forall x)(\forall f)(\forall f')(f, f' \text{ are attempts} \wedge x \in \operatorname{dom} f \cap \operatorname{dom} f' \Rightarrow f(x) = f'(x))$$

by $\in$-induction: if $f(y) = f'(y)$ for all $y \in x$, then $f(x) = f'(x)$. Also,

$$(\forall x)(\exists f)(f \text{ is an attempt} \wedge x \in \operatorname{dom} f)$$

by $\in$-induction. Indeed, if for all $y \in x$ there exists an attempt defined at y, then for each $y \in x$ there is a unique attempt f_y defined on $TC(\{y\})$. Let $f = \bigcup\{f_y \mid y \in x\}$, which is an attempt with domain $TC(x)$. We can then define $f' = f \cup \{(x, G(f|_x))\}$. This is an attempt defined at x. We can then take $q(x, y)$ to be

$$(\exists f)(f \text{ is an attempt} \wedge x \in \operatorname{dom} f \wedge f(x) = y)$$

This defines the function-class F as required. Uniqueness follows from the fact that if F, F' are suitable function-classes, we have $(\forall x)(F(x) = F'(x))$ by $\in$-induction. □

5.5. Well-founded relations

Note the similarity between the proofs of $\in$-induction and $\in$-recursion and the proofs of induction and recursion on ordinals. These proofs are not specific to the relation $\in$; we only used some of its properties.

(i) p is *well-founded*: every nonempty set has a p-minimal element.

(ii) p is *local*: $\{x \mid p(x, y)\}$ is a set. This was required to build the p-transitive closure.

Therefore, p-induction and p-recursion hold for all relation-classes p that are well-founded and local. In particular, if r is a well-founded relation on a set a, it is clearly local and hence we have r-induction and r-recursion. The theorems about induction and recursion on ordinals are therefore special cases of this, as a well-ordering is precisely a well-founded total order.

On the set $\{a, b, c\}$, let r be the relation arb, brc. Choosing $a' = \varnothing, b' = \{\varnothing\}, c' = \{\{\varnothing\}\}$, the map $f : \{a, b, c\} \to \{a', b', c'\}$ given by $x \mapsto x'$ is a bijection with a transitive set such that xry if and only if $f(x) \in f(y)$. This models the relation r by $\in$.

We say that a relation r on a set a is *extensional* if

$$(\forall x \in a)(\forall y \in a)((\forall z \in a)(zrx \Leftrightarrow zry) \Rightarrow x = y)$$

The relation r in the above example is extensional.

Theorem (Mostowski's collapsing theorem)**.** Let r be a relation on a set a that is well-founded and extensional. Then, there exists a transitive set b and a bijection $f : a \to b$ such that

$$(\forall x \in a)(\forall y \in a)(xry \Leftrightarrow f(x) \in f(y))$$

Moreover, b and f are unique.

This is an analogue of subset collapse from the section on ordinals. Transitive sets are playing the role of initial segments. Note that the well-foundedness and extensionality conditions are clearly necessary for the theorem, consider $(\mathbb{Z}, <)$ or $(\{a, b, c, \}, <)$ with $a < b, a < c$ for counterexamples.

Proof. We define the function f by $f(x) = \{f(y) \mid yrx\}$ using r-recursion. Note that f is a function by the axiom of replacement as it is given by a function-class F obtained from r-recursion that is defined on the set a. Let $b = \{f(x) \mid x \in a\}$; this is a set by the axiom of replacement. Clearly f is surjective by the definition of b, and b is transitive by definition.

We claim that f is injective, and then we have that yrx if and only if $f(y) \in f(x)$ by definition of f. We show

$$(\forall x \in a)(\forall x' \in a)(f(x') = f(x) \Rightarrow x' = x)$$

by r-induction on x. Suppose that $(\forall yrx)(\forall z \in a)(f(y) = f(z) \Rightarrow y = z)$, we have $f(x) = f(x')$, and we want to show that $x = x'$. Note that $\{f(y) \mid yrx\} = \{f(z) \mid zrx'\}$

by the definition of f as $f(x) = f(x')$. So $\{y \mid yrx\} = \{z \mid zrx'\}$, so $x = x'$ as r is extensional. Uniqueness holds by r-induction, as we must have $f(x) = \{f(y) \mid yrx\}$ for all $x \in a$. □

In particular, every well-ordered set has a unique order isomorphism to a unique transitive set well-ordered by $\in$. We can now define that an ordinal is a transitive set well-ordered by $\in$ (or equivalently, totally-ordered, due to the axiom of foundation). For example, $\varnothing$ is an ordinal, $n \in \omega$ is an ordinal, ω is also an ordinal, and so on. Therefore, each well-ordering is order-isomorphic to a unique ordinal called its order type, by Mostowski collapse.

Remark. If x, y are elements of a well-ordered set a with $y < x$, then the order type of I_x, which is precisely the image $f(x)$ under the Mostowski collapse, has an element $f(y)$, the order type of I_y. In particular, given two ordinals α, β, the statement $\alpha < \beta$ is equivalent to $\alpha \in \beta$. Hence $\alpha = \{\beta \mid \beta < \alpha\}$. Thus, $\alpha^+ = \alpha \cup \{\alpha\}$, and $\sup\{\alpha_i \mid i \in I\} = \bigcup\{\alpha_i \mid i \in I\}$.

5.6. The universe of sets

We would like the universe to be V-shaped, in the sense that we begin with $\varnothing$ and continue taking power sets to create larger and larger sets. Define sets V_α for each ordinal α by

- $V_0 = \varnothing$;
- $V_{\alpha+1} = \mathcal{P}(V_\alpha)$;
- $V_\lambda = \bigcup\{V_\alpha \mid \alpha < \lambda\}$ for a nonzero limit ordinal λ.

This can be viewed as a well-founded recursion on ordinals, or $\in$-recursion on the universe but mapping non-ordinals to $\varnothing$. For example, $V_\omega = V_0 \cup V_1 \cup \dots$, and $V_{\omega+1} = \mathcal{P}(V_\omega)$. We will now show that every set is contained within some V_α.

Lemma. Each V_α is transitive.

Proof. We show this by induction on α. Clearly $V_0 = \varnothing$ is transitive. Suppose V_α is transitive. Then $V_{\alpha+1}$ is transitive as the power set of a transitive set is transitive. Indeed, if x is transitive and $z \in y \in \mathcal{P}(x)$, we have $z \in x$, so $z \subseteq x$ as x is transitive, so $z \in \mathcal{P}(x)$. Now suppose λ is a limit ordinal, and that the V_α are transitive for $\alpha < \lambda$. Any union of transitive sets is transitive, so V_λ is transitive. □

Lemma. Let $\alpha \leq \beta$. Then $V_\alpha \subseteq V_\beta$.

Proof. We show this by induction on β for a fixed α. If $\beta = \alpha$, $V_\alpha \subseteq V_\beta$ is trivial. For successors, note that $V_\beta \subseteq \mathcal{P}(V_\beta)$ as V_β is transitive. So if $V_\alpha \subseteq V_\beta$, then $V_\alpha \subseteq V_{\beta+1}$. Limits are trivial. □

Theorem. Every set x belongs to V_α for some α.

If we could construct the set V defined as the union of the V_α over all ordinals α, V would be a model of ZF.

Remark. Note that $x \subseteq V_\alpha$ if and only if $x \in V_{\alpha+1}$, so it suffices to show that each set x is a subset of some V_α. Once we have $x \subseteq V_\alpha$ for some α, there is a least such α, called the *rank* of x. For example, the rank of $\varnothing$ is 0, the rank of 1 is 1, the rank of ω is ω, and in general the rank of any ordinal α is α. Intuitively, the rank of a set is the time at which it was created.

Proof. We proceed by $\in$-induction on x; we may assume that for all $y \in x$, there exists α such that $y \subseteq V_\alpha$, so $y \subseteq V_{\mathrm{rank}(y)}$. Thus, for each $y \in x$, $y \in V_{\mathrm{rank}(y)+1}$, so define $\alpha = \sup\{\mathrm{rank}(y) + 1 \mid y \in x\}$. Then for all $y \in x$, we have $y \in V_\alpha$. So $x \subseteq V_\alpha$ as required. □

The ordinals can be viewed as the backbone of the universe of sets; each V_α can be thought of as resting on the ordinal α.

Remark. The V_α are called the *von Neumann hierarchy*. The above proof shows that for all x, $\mathrm{rank}(x) = \sup\{\mathrm{rank}(y) + 1 \mid y \in x\}$. For example, the rank of $\{\{2, 3\}, 6\}$ is

$$\sup\{\mathrm{rank}\{2, 3\} + 1, 6 + 1\} = \sup\{5, 7\} = 7$$

6. Cardinals

6.1. Definitions

We will study the possible sizes of sets in ZFC. Write $x \leftrightarrow y$ if there exists a bijection from x to y; we wish to define $\operatorname{card}(x) = |x|$ such that $x \leftrightarrow y$ if and only if $\operatorname{card}(x) = \operatorname{card}(y)$. This cannot be formulated as an equivalence class, due to Russell's paradox. However, for any x, there exists an ordinal α such that $x \leftrightarrow \alpha$ by the well-ordering theorem. Hence, we can define $\operatorname{card}(x)$ to be the least ordinal that x bijects with. We say that a set m is a *cardinality* or a *cardinal* if $m = \operatorname{card}(x)$ for some set x.

If we were studying sets in ZF and not ZFC, there may not be an ordinal that bijects with a given set x. However, we can apply *Scott's trick*, which is as follows. We can consider the least α such that there exists $y \leftrightarrow x$ with $\operatorname{rank}(y) = \alpha$. This is often called the *essential rank* of x. In this case, we let $\operatorname{card}(x)$ be the set $\{y \subseteq V_\alpha \mid y \leftrightarrow x\}$.

6.2. The hierarchy of alephs

An ordinal is *initial* if it does not biject with any smaller ordinal. Any finite ordinal is initial, and ω, ω_1 are initial. For any set x, $\gamma(x)$ is initial. ω^2 is not initial as it bijects with ω. We define ω_α for each ordinal α by recursion.

- $\omega_0 = \omega$;
- $\omega_{\alpha+1} = \gamma(\omega_\alpha)$;
- $\omega_\lambda = \sup\{\omega_\alpha \mid \alpha < \lambda\}$ for a nonzero limit ordinal λ.

Each of these ordinals is initial, and every initial ordinal β is of the form ω_α. Indeed, the ω_α are unbounded, as $\omega_\alpha \geq \alpha$ for each α by induction, so there exists a least ordinal δ such that $\beta < \omega_\delta$. δ must be a successor, otherwise $\omega_\delta = \sup\{\omega_\alpha \mid \alpha < \beta\}$, contradicting the definition of δ. So $\delta = \alpha + 1$, so $\omega_\alpha \leq \beta < \omega_{\alpha+1}$. Hence $\beta = \omega_\alpha$, otherwise we contradict $\omega_{\alpha+1} = \gamma(\omega_\alpha)$.

Since we have potentially different definitions of cardinals, we will write $\aleph_\alpha$ for $\operatorname{card}(\omega_\alpha)$ to avoid ambiguity. The $\aleph_\alpha$ are precisely the cardinalities of the infinite sets. In ZF without AC, the $\aleph_\alpha$ are the cardinalities of the well-orderable sets.

For cardinals m, n, we write $m \leq n$ if there exists an injection from M to N where $\operatorname{card}(M) = m, \operatorname{card}(N) = n$. Similarly, we write $m < n$ if $m \leq n$ and $m \neq n$. For example, $\operatorname{card}(\omega) < \operatorname{card}(\mathcal{P}(\omega))$. By the Schröder–Bernstein theorem, if $m \leq n$ and $n \leq m$, then $m = n$. Hence, $\leq$ is a partial order on cardinals. This is in fact a total order in ZFC, since we can well-order the two sets in question, and one injects into the other; alternatively, the $\aleph$ numbers are clearly totally ordered.

6.3. Cardinal arithmetic

Let m, n be cardinals. Then,

(i) $m + n = \operatorname{card}(M \amalg N)$;

(ii) $m \cdot n = \operatorname{card}(M \times N)$;

(iii) $m^n = \operatorname{card}(M^N)$;

where $m = \operatorname{card}(M)$, $n = \operatorname{card}(N)$, and M^N is the set of functions $N \to M$. The choice of representatives M, N do not influence the result. We can also define $\sum_{i\in I} m_i = \operatorname{card}\big(\coprod_{i\in I} M_i\big)$; this is only well-defined assuming the axiom of choice, as forming the bijection requires infinitely many choices.

Example. $\mathbb{R}, \mathcal{P}(\omega), \{0,1\}^\omega$ biject. Hence, $\operatorname{card}(\mathbb{R}) = \operatorname{card}(\mathcal{P}(\omega)) = 2^{\aleph_0}$. In particular, cardinal exponentiation and ordinal exponentiation do not coincide, as $2^\omega = \omega$.

The cardinality of the set of sequences of reals is

$$\operatorname{card}(\mathbb{R}^\omega) = (2^{\aleph_0})^{\aleph_0} = 2^{\aleph_0 \cdot \aleph_0} = 2^{\aleph_0}$$

Note that this statement requires that addition and multiplication are commutative, $\aleph_0 \cdot \aleph_0 = \aleph_0$ as $\omega \times \omega$ bijects with ω, and that $(m^n)^p = m^{np}$. The latter holds as $(M^N)^P$ is the set of functions $P \to (N \to M)$, and $M^{N\times P}$ is the set of functions $N \times P \to M$.

Theorem. $m^2 = m$ for all infinite cardinals m.

Proof. We show by induction that $\aleph_\alpha^2 = \aleph_\alpha$ for all α. Define a well-ordering of $\omega_\alpha \times \omega_\alpha$ by 'going up in squares':

$$\begin{aligned}(x,y) < (z,w) \iff &(\max(x,y) < \max(z,w)) \vee \\ &(\max(x,y) = \max(z,w) = \beta \\ &\quad \wedge (y < \beta, z < \beta \vee x = z = \beta, y < w \vee y = w = \beta, x < z))\end{aligned}$$

For any $\delta \in \omega_\alpha \times \omega_\alpha$, $\delta \in \beta \times \beta$ for some $\beta < \omega_\alpha$, as ω_α is a limit ordinal. By induction, we can assume $\beta \times \beta$ bijects with β (or β is finite). Hence, the initial segment I_δ is contained in $\beta \times \beta$ and hence has cardinality at most $\operatorname{card}(\beta \times \beta) < \operatorname{card}(\omega_\alpha)$.

Therefore, the well-ordering has order type at most ω_α. Thus, $\omega_\alpha \times \omega_\alpha$ injects into ω_α, and the converse injection is trivial. So $\omega_\alpha \times \omega_\alpha$ bijects with ω_α. □

Corollary. For any ordinals $\alpha < \beta$, we have $\aleph_\alpha + \aleph_\beta = \aleph_\alpha \cdot \aleph_\beta = \aleph_\beta$.

Proof.

$$\aleph_\beta \leq \aleph_\alpha + \aleph_\beta \leq 2 \cdot \aleph_\beta \leq \aleph_\alpha \aleph_\beta \leq \aleph_\beta^2 = \aleph_\beta$$

□

Hence, for example, $X \amalg X$ bijects with X for any infinite set X.

Cardinal exponentiation is not as simple as addition and multiplication. For instance, in ZF, $2^{\aleph_0}$ need not even be an aleph number, for instance if $\mathbb{R}$ is not well-orderable. In ZFC, the statement $2^{\aleph_0} = \aleph_1$ is independent of the axioms; this is called the *continuum hypothesis*. ZFC does not even decide if $2^{\aleph_0} < 2^{\aleph_1}$. Even today, not all implications about cardinal exponentiation (such as $\aleph_\alpha^{\aleph_\beta}$) are known.

7. Incompleteness

We aim to show that PA is incomplete: there is a sentence p such that PA does not prove p or $\neg p$. Equivalently, there is a sentence p that is true in $\mathbb{N}$ but PA $\nvdash p$. In this section, by 'true' we mean true in $\mathbb{N}$, and by 'unprovable' we mean PA does not prove it, so more concisely we wish to find an unprovable true sentence. Our aim is to find a sentence p that asserts that it is not provable in PA; then p is true if and only if p is not provable. Then the proof is complete, as if p is false, p is provable and hence true by soundness.

7.1. Definability

Recall that a subset $S \subseteq \mathbb{N}$ is *definable* if there is a formula p with free variable x such that $m \in S$ if and only if $p(m)$ is true. For example, the set of primes is definable, taking $p(x)$ to be $(x \neq 1) \land (\forall y)(\forall z)(yz = x \Rightarrow (y = 1) \lor (z = 1))$. We might say that '$m$ is prime' is definable.

A function $f : \mathbb{N} \to \mathbb{N}$ is similarly called definable if there is a formula p with free variables x, y such that $f(m) = n$ if and only if $p(m, n)$ is true. The function $f(x) = \left\lfloor \frac{x}{2} \right\rfloor$ is definable, setting $p(x, y)$ to be $(x = 2y) \lor (x = 2y + 1)$. Similarly, x^2 is definable. In fact, any function f given by an algorithm is definable in PA, but this will not be proven in this section.

7.2. Coding

L has symbols

$$0, s, +, \cdot, =, \bot, \Rightarrow, (,), \forall, x, '$$

labelling each variable x, x', x'' and so on. We code each symbol by assigning it a number, so $v(0) = 1, \dots, v(') = 12$. A formula p is encoded by

$$c(p) = 2^{v(\text{first symbol})} 3^{v(\text{second symbol})} \dots n\text{th prime}^{v(n\text{th symbol})}$$

For instance, if p is the assertion $(\forall x)(x = 0)$, then

$$c(p) = 2^8 3^{10} 5^{11} 7^9 11^8 13^{11} 17^5 19^1 23^9$$

Clearly, not all numbers encode formulae. We will write S_n for the formula encoded by n, with $S_n = \bot$ if n does not encode a formula. Observe that the statement 'n codes a formula' is definable, as there is an algorithm to decide it.

The statement 'l, m, n code formulae and S_n is obtained from S_l, S_m by modus ponens' is definable. The analogous statement for generalisation is also definable in a similar way. The axioms of PA are clearly definable, and 'n codes a logical axiom or axiom of PA' is definable. Given formulae $p_1, \dots, p_n$, we code the sequence as

$$s(p_1, \dots, p_n) = 2^{c(p_1)} 3^{c(p_2)} \dots n\text{th prime}^{c(p_n)}$$

Thus, 'n codes a proof' is definable, and 'n codes a proof of S_m' is definable. Let $\theta(m, n)$ be a formula defining 'n codes a proof of S_m'. Let $\phi(m)$ = 'S_m is provable' is definable, as $\phi(m) = (\exists n)(\theta(m, n))$.

7.3. Gödel's incompleteness theorem

Consider $\chi(m)$ = 'm codes a formula S_m with one free variable, and $S_m(m)$ is unprovable'. This is definable, so is given by some formula $p(x)$, so $\chi(m)$ holds if and only if $p(m)$ holds. Let N be the code for $p(x)$. Then, $p(N)$ is the assertion that N codes a formula S_N with one free variable, such that $S_N(N)$ is unprovable. Note that $S_N = p$ and $S_N(N) = p(N)$, so $p(N)$ asserts that $p(N)$ is unprovable. The sentence $p(N)$ suffices for the above argument, so we have shown the following theorem.

Theorem. PA is incomplete.

Note that if our proof above could be written in PA, we would then have that $p(N)$ is provable in PA. One can check that the proof used the fact that a model of PA exists (namely, $\mathbb{N}$, although this was not particularly important). We thus used the statement Con(PA), that PA is consistent, or equivalently,

$$(\forall x)(x \text{ does not code a proof of } \bot)$$

Thus, our proof above formalises to the statement

$$\mathsf{PA} \cup \{\mathrm{Con}(\mathsf{PA})\} \vdash p(N)$$

The next theorem then follows.

Theorem. PA $\nvdash$ Con(PA).

PA is incomplete, but we cannot add any true sentence t to obtain a complete theory. Indeed, the proof above can be performed on this new theory PA $\cup \{t\}$ to show that it is incomplete. However, PA can certainly be extended to some complete theory by taking the set of all sentences that hold in $\mathbb{N}$. We cannot use the above proof to show that T is incomplete, since this would immediately derive a contradiction. Almost all of the above proof is still valid, so the only invalid part must lead to this contradiction; there must be no algorithm to decide truth of sentences in PA.

Theorem. T is not decidable.

Note that ZFC $\vdash$ Con(PA), where Con(PA) represents the sentence

$$(\forall x \in \omega)(x \text{ does not code a proof of } \bot)$$

This is because ZFC proves that PA has a model, namely ω. However, as for the above theorems, we obtain the following.

Theorem. ZFC is incomplete (if ZFC is consistent).

Theorem. ZFC $\nvdash$ Con(ZFC) (if ZFC is consistent).