Studies in Algebra and Group Theory

S. D. V. Stephens

Notes on Harvard University's Math 55a,
taught by Denis Auroux in the fall of 2025

Contents

I Sets and Groups

1 Set Theory **3**
- 1.1 Ordered Pairs . . . 3
- 1.2 Relations . . . 4
- 1.3 Functions . . . 8
- 1.4 Families . . . 11
- 1.5 Inverses and Composites . . . 13
- 1.6 Quotient Sets and Fibers . . . 16
- 1.7 Cardinality . . . 19

2 Groups and Homomorphisms **25**
- 2.1 The Group Axioms . . . 25
- 2.2 Fundamental Examples . . . 28
- 2.3 Constructions: Products and Sums . . . 34
- 2.4 Subgroups . . . 36
- 2.5 Homomorphisms . . . 41
- 2.6 Orders of Elements and Cyclic Groups . . . 46

3 Quotients, Actions, and Presentations **53**
- 3.1 Cosets and Quotient Groups . . . 53
- 3.2 Group Actions . . . 61
- 3.3 The Isomorphism Theorems . . . 68
- 3.4 The Symmetric Group . . . 74
- 3.5 Exact Sequences. . . . 83
- 3.6 Free Groups and Presentations . . . 89

II Linear Algebra

4 Rings, Fields, and Vector Spaces **103**
- 4.1 Rings and Fields . . . 103
- 4.2 Ring and Field Homomorphisms . . . 107

4.3 Polynomial Rings and Field Extensions . . . 108
4.4 The Characteristic of a Field . . . 114
4.5 Vector Spaces . . . 116
4.6 Span, Linear Independence, and Bases . . . 125
4.7 Finite-Dimensional Vector Spaces and Dimension . . . 129

5 Linear Maps, Matrices, and Determinants 139

5.1 Linear Maps . . . 139
5.2 Matrices of Linear Maps . . . 143
5.3 Change of Basis . . . 149
5.4 Determinants . . . 153

6 Kernels, Images, and Duality 167

6.1 Kernel and Image . . . 167
6.2 Duality . . . 178

7 Quotient Spaces, Eigenvalues, and Jordan Form 189

7.1 Quotient Spaces . . . 189
7.2 Polynomials in an Operator . . . 193
7.3 Invariant Subspaces . . . 195
7.4 Eigenvalues and Eigenvectors . . . 197
7.5 Existence of Eigenvectors and Triangular Form . . . 205
7.6 Generalized Eigenspaces . . . 208
7.7 Nilpotent Operators . . . 215
7.8 Jordan Normal Form . . . 221
7.9 The Cayley–Hamilton Theorem and the Minimal Polynomial . . . 226
7.10 Real Vector Spaces and Complexification . . . 233

8 Categories: An Interlude 239

8.1 The Language of Categories . . . 239
8.2 Functors and Natural Transformations . . . 245
8.3 Universal Properties . . . 250

9 Inner Products, Bilinear Forms, and Tensors 257

9.1 Bilinear Forms and Inner Products . . . 257
9.2 Orthonormal Bases . . . 270
9.3 The Adjoint . . . 275
9.4 Self-Adjoint and Normal Operators . . . 279
9.5 Orthogonal and Unitary Operators . . . 285
9.6 Bilinear and Quadratic Forms . . . 296
9.7 Symplectic Forms . . . 302
9.8 Tensor Products . . . 308
9.9 Multilinear Maps and the Tensor Algebra . . . 316
9.10 Symmetric and Exterior Powers . . . 319
9.11 Determinants, Trace, and Volume Forms . . . 331
9.12 Interlude: Tensors in Differential Geometry and Physics . . . 341

III The Structure of Groups

10 Modules and Finitely Generated Abelian Groups 349
10.1 Modules over a Ring . . . 349
10.2 Classification of Finitely Generated Abelian Groups . . . 356

11 Group Actions, Conjugacy, and the Sylow Theorems 367
11.1 Orbits, Stabilizers, and Burnside's Lemma . . . 367
11.2 Conjugation and the Class Equation . . . 370
11.3 Finite Subgroups of $SO(3)$ . . . 373
11.4 Conjugacy in S_n and A_n . . . 380
11.5 The Sylow Theorems . . . 385
11.6 Semidirect Products . . . 389

12 Generators, Presentations, and Cayley Graphs 397
12.1 Cayley Graphs . . . 397
12.2 Permutation Diagrams and Normal Forms in S_n . . . 401
12.3 The Braid Group . . . 405
12.4 $SL_2(\mathbb{Z})$ and $PSL_2(\mathbb{Z})$ . . . 408

IV Representation Theory

13 Representations and Characters 419
13.1 Representations and Subrepresentations . . . 419
13.2 Maschke's Theorem and Complete Reducibility . . . 424
13.3 Schur's Lemma . . . 426
13.4 Characters . . . 429
13.5 Orthogonality of Characters . . . 433
13.6 Characters Form an Orthonormal Basis. . . . 437

14 Character Tables and the Representation Ring 443
14.1 The Character Table of S_4 . . . 444
14.2 The Character Table of A_4. . . . 446
14.3 The Character Table of S_5 . . . 447
14.4 The Character Table of A_5. . . . 449
14.5 The Representation Ring . . . 452

15 Restriction and Induction 457
15.1 Restriction . . . 457
15.2 Induced Representations . . . 458
15.3 Frobenius Reciprocity. . . . 462
15.4 Irreducibility of Induced Representations . . . 464
15.5 Example: $S_4 \supset S_3$ . . . 466

16 Real and Quaternionic Representations **471**
16.1 Representations over $\mathbb{R}$ . . . 471
16.2 Complexification . . . 472
16.3 Characterizing Real Representations . . . 474
16.4 Invariant Bilinear Forms . . . 474
16.5 Quaternionic Representations . . . 476
16.6 The Frobenius–Schur Indicator . . . 478
16.7 Examples . . . 480

V Further Topics

17 Further Topics in Algebra **485**
17.1 Commutative Rings . . . 485
17.2 Galois Theory . . . 495
17.3 Lie Groups and Lie Algebras . . . 504

Conventions

Numbering. Definitions, theorems, propositions, lemmas, corollaries, examples and exercises are numbered together within each section: Theorem 3.2.4 is the fourth numbered item of Section 3.2.

Maps and permutations. Maps are written on the left of their arguments and compose from right to left: $g \circ f$ means first f, then g. Products of permutations follow the same rule, so $\sigma\tau$ applies τ first.

Fields and vector spaces. The letter k denotes a field. A statement that needs a vector space to be finite-dimensional says so.

Duals and adjoints. The dual of a vector space V is $V^* = \mathrm{Hom}(V, k)$. The dual (or transpose) of a linear map $\varphi\colon V \to W$ is $\varphi^t\colon W^* \to V^*$, $\varphi^t(\ell) = \ell \circ \varphi$; in dual bases its matrix is the transpose A^t of the matrix of φ. The symbol φ^* is reserved for the adjoint of a map between inner product spaces, and $A^* = \overline{A}^{\,t}$ for the conjugate transpose of a matrix.

Hermitian forms. Hermitian forms and inner products are conjugate-linear in the first variable and linear in the second: $\langle \lambda u, v\rangle = \overline{\lambda}\,\langle u, v\rangle$ and $\langle u, \lambda v\rangle = \lambda\,\langle u, v\rangle$. The standard inner product on $\mathbb{C}^n$ is $\langle z, w\rangle = \sum_i \overline{z_i}\, w_i = z^* w$.

Group actions. Group actions are left actions unless stated otherwise.

PART I

Sets and Groups

CHAPTER 1

Set Theory

This chapter collects the set theory that the rest of the book takes for granted: ordered pairs, relations, functions, families of sets, inverse images and composites, equivalence relations and the quotient sets they produce, and cardinality. Much of it follows Halmos's *Naive Set Theory*. Throughout, $\mathcal{P}(X)$ denotes the ***power set*** of a set X, the set of all subsets of X; $X - A$ denotes the set of elements of X that do not lie in A; and $\mathbb{N} = \{0, 1, 2, \ldots\}$.

1.1 Ordered Pairs

We begin with a discussion of ordered pairs. A set carries no order: $\{1, 2, 3, 4\} = \{1, 3, 2, 4\} = \{3, 4, 2, 1\}$, and so on. Say we want to list the elements of $\{1, 2, 3, 4\}$ in the order $3, 2, 1, 4$, for some arbitrary reason; we may then consider the collection

$$\mathfrak{C} = \{\{3\}, \{3, 2\}, \{3, 2, 1\}, \{3, 2, 1, 4\}\}.$$

The order is recorded by the number of sets of the collection in which each element appears: 3 lies in all four, 2 in three of them, 1 in two, and 4 in one.

Applying the same idea to a collection of two sets built from elements a and b gives the definition of a 2-tuple, or ***ordered pair***:

$$(a, b) \coloneqq \{\{a\}, \{a, b\}\} \quad \text{and} \quad (b, a) \coloneqq \{\{b\}, \{b, a\}\}.$$

> NOTE This way of encoding ordered pairs as sets is called the *Kuratowski* definition of the ordered pair.

The one property of ordered pairs that the rest of the book uses is that a pair determines its coordinates.

Proposition 1.1.1 (Equality of ordered pairs)
For all a, b, c, d, we have $(a, b) = (c, d)$ if and only if $a = c$ and $b = d$.

Proof. If $a = c$ and $b = d$, the two sets are equal. Conversely, suppose $\{\{a\},\{a,b\}\} = \{\{c\},\{c,d\}\}$. If $a = b$, the left side is $\{\{a\}\}$, which has one element, so the right side has one element too; hence $\{c\} = \{c,d\}$, which forces $c = d$, and then $\{a\} = \{c\}$ gives $a = c$. Thus $a = b = c = d$. If $a \neq b$, the left side has two elements, the one-element set $\{a\}$ and the two-element set $\{a,b\}$. So the right side also has two elements, which forces $c \neq d$, and matching the one-element members gives $\{a\} = \{c\}$, so $a = c$. Matching the two-element members gives $\{a,b\} = \{a,d\}$; since $b \in \{a,d\}$ and $b \neq a$, we get $b = d$. □

Ordered triples and, more generally, n-tuples can be built the same way; Section 1.4 gives a uniform description of them as families.

Definition 1.1.2 (Cartesian product)

The ***Cartesian product*** of sets A and B is the set of ordered pairs

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

We write A^2 for $A \times A$.

Since $\{a\}$ and $\{a,b\}$ are subsets of $A \cup B$, every pair (a,b) with $a \in A$ and $b \in B$ is a set of subsets of $A \cup B$, that is, $(a,b) \in \mathcal{P}(\mathcal{P}(A \cup B))$. So $A \times B$ is a subset of $\mathcal{P}(\mathcal{P}(A \cup B))$; in particular, it is a set.

1.2 Relations

Our review of ordered pairs leads us to relations. Informally, a relation is a property that a pair of elements may or may not have. Think of equality, which holds for the pair $(2+3, 1+4)$; or, for something more interesting, of the set of all (x,y) such that x is a man, y is an enby (slay), and x is married to y (many such cases). Formally, we identify a relation with the set of pairs for which it holds.

Definition 1.2.1 (Relations)

A ***relation*** between sets A and B is a subset $R \subseteq A \times B$; we write xRy for $(x,y) \in R$. A relation ***on*** a set X is a relation between X and X, that is, a subset $R \subseteq X \times X$. The ***domain*** and ***range*** of R are

$$\operatorname{dom} R = \{x \mid xRy \text{ for some } y\}, \qquad \operatorname{ran} R = \{y \mid xRy \text{ for some } x\}.$$

Consider the following examples:

(a) $R \coloneqq \{(x, y) \in \mathbb{N}^2 \mid y - x > 0\}$, so that xRy means $x < y$: the relation *less than*;
(b) $R \coloneqq \{(a, B) \in A \times \mathcal{P}(A) \mid a \in B\}$, so that aRB means $a \in B$: the relation of *belonging* (being an element of);
(c) $R \coloneqq \{(c, d) \in \{\text{all cats}\} \times \{\text{all dogs}\} \mid c \text{ and } d \text{ live in the same house}\}$.

Naturally, there are infinitely many possible relations, but one class of them concerns us for now: the relations that declare elements equivalent to one another, in a sense that the next definition makes precise.

Definition 1.2.2 (Equivalence relations)

Let R be a relation on a set X. We say that R is

(a) ***reflexive*** if xRx for every $x \in X$;
(b) ***symmetric*** if xRy implies yRx;
(c) ***transitive*** if xRy and yRz imply xRz.

An ***equivalence relation*** on X is a relation on X that is reflexive, symmetric, and transitive. Equivalence relations are usually denoted $\sim$.

Question 1.2.3

The three properties in the definition are independent of one another. For each of them, find an explicit relation that has that property and lacks the other two.

Solution. For reflexivity, let

$$R \coloneqq \{(1,1), (2,2), (3,3), (1,2), (2,3)\} \subseteq \{1,2,3\}^2.$$

Since $(1,1), (2,2), (3,3) \in R$, the relation R is reflexive on $\{1,2,3\}$. However, $(1,2) \in R$ but $(2,1) \notin R$, so R is not symmetric. For transitivity, we have $(1,2)$ and $(2,3)$ in R, so we can set up the first step, but $(1,3) \notin R$; thus transitivity fails. This example is a reminder that we need not think of concrete properties that can be seen in front of our eyes (shapes, dogs, even equality!) in order to have a mathematical relation: any set of pairs will do. For symmetry, let X be a set with at least two elements and consider $R \coloneqq \{(a, b) \in X^2 \mid a \neq b\}$. No element satisfies aRa, so R is not reflexive. It is symmetric, since $a \neq b$ implies $b \neq a$. Transitivity may seem plausible at first, but for $a \neq b$ we have aRb and bRa, and transitivity would then demand aRa, which is false.

For transitivity, consider $x > y$ on $\mathbb{N}$. It is not reflexive, since x is never greater than itself, and it is not symmetric, since $x > y$ never implies $y > x$. It is transitive: $x > y$ and $y > z$ imply $x > z$.

1.2.1 Equivalence Classes and Partitions

An equivalence relation on X declares certain elements of X equivalent to one another. Gathering mutually equivalent elements together produces a smaller set, whose elements are the resulting classes.

Definition 1.2.4 (Equivalence classes)

Let $\sim$ be an equivalence relation on a set X. The ***equivalence class*** of $a \in X$ is

$$[a] = \{x \in X \mid a \sim x\}.$$

Any element of $[a]$ is called a ***representative*** of the class.

By reflexivity, $a \in [a]$.

Lemma 1.2.5 (Equal classes)

Let $\sim$ be an equivalence relation on X, and let $a, b \in X$.

(a) $[a] = [b]$ if and only if $a \sim b$.

(b) If $[a] \cap [b] \neq \emptyset$, then $[a] = [b]$.

Proof. (a) Suppose $a \sim b$. If $x \in [b]$, then $b \sim x$, and transitivity gives $a \sim x$; so $[b] \subseteq [a]$. By symmetry $b \sim a$, and the same argument gives $[a] \subseteq [b]$. Conversely, if $[a] = [b]$, then $b \in [b] = [a]$, that is, $a \sim b$.

(b) If $c \in [a] \cap [b]$, then $a \sim c$ and $b \sim c$; by symmetry $c \sim b$, and by transitivity $a \sim b$. Now apply (a). □

Definition 1.2.6 (Partitions)

A ***partition*** of a set X is a collection $\mathcal{C}$ of nonempty subsets of X such that every element of X belongs to exactly one member of $\mathcal{C}$; equivalently, the members of $\mathcal{C}$ are nonempty and pairwise disjoint, and their union is X. The relation ***induced*** by $\mathcal{C}$ is the relation $\sim_{\mathcal{C}}$ on X given by

$$x \sim_{\mathcal{C}} y \iff \text{some } A \in \mathcal{C} \text{ contains both } x \text{ and } y.$$

Definition 1.2.7 (Quotient sets)

Let $\sim$ be an equivalence relation on X. The ***quotient set*** of X by $\sim$ is the set of its equivalence classes,

$$X/\sim = \{[a] \mid a \in X\} \subseteq \mathcal{P}(X).$$

Equivalence relations and partitions turn out to be two descriptions of the same thing.

Proposition 1.2.8 (Equivalence relations and partitions)

Let X be a set.

(a) If $\sim$ is an equivalence relation on X, then $X/\sim$ is a partition of X, and the relation it induces is $\sim$.

(b) If $\mathfrak{C}$ is a partition of X, then $\sim_{\mathfrak{C}}$ is an equivalence relation on X, and its set of equivalence classes is exactly $\mathfrak{C}$: $X/\sim_{\mathfrak{C}} = \mathfrak{C}$.

Proof. (a) Each class $[a]$ is nonempty, since it contains a, and each $x \in X$ lies in the class $[x]$. If x lies in two classes $[a]$ and $[b]$, they intersect, so $[a] = [b]$ by Lemma 1.2.5. Hence $X/\sim$ is a partition, and the only class containing x is $[x]$. Consequently some class contains both x and y if and only if $[x] = [y]$, which by Lemma 1.2.5 happens if and only if $x \sim y$.

(b) For $x \in X$, let A_x be the unique member of $\mathfrak{C}$ containing x. If $x, y \in A$ for some $A \in \mathfrak{C}$, then $A = A_x$ by uniqueness; so $x \sim_{\mathfrak{C}} y$ if and only if $y \in A_x$. Now $\sim_{\mathfrak{C}}$ is reflexive, since $x \in A_x$; it is symmetric, since its definition is symmetric in x and y; and it is transitive, since $x \sim_{\mathfrak{C}} y$ and $y \sim_{\mathfrak{C}} z$ give $y \in A_x$, hence $A_y = A_x$, and $z \in A_y = A_x$. The class of x is $\{y \mid y \in A_x\} = A_x$, so $X/\sim_{\mathfrak{C}} = \{A_x \mid x \in X\}$. Every member A of $\mathfrak{C}$ is nonempty, and $A = A_x$ for any $x \in A$; hence $X/\sim_{\mathfrak{C}} = \mathfrak{C}$. □

Example 1.2.9 (Congruence mod n)

Fix an integer $n \geqslant 1$, and for $a, b \in \mathbb{Z}$ set $a \sim b$ if n divides $b - a$. This is an equivalence relation, called ***congruence mod*** n and also written $a \equiv b \pmod{n}$: n divides $a - a = 0$; if n divides $b - a$, it divides $a - b$; and if n divides $b - a$ and $c - b$, it divides their sum $c - a$. The class of a is $[a] = \{a + mn \mid m \in \mathbb{Z}\}$. By division with remainder, every integer can be written as $a = qn + r$ with $q \in \mathbb{Z}$ and $0 \leqslant r \leqslant n - 1$, so a is congruent to its remainder r. No two of $0, 1, \ldots, n - 1$ are congruent, since their differences have absolute value less than n. So every integer is congruent to exactly one

of $0, 1, \ldots, n-1$, and the quotient set has exactly n elements:

$$\mathbb{Z}/\sim = \{[0], [1], \ldots, [n-1]\}.$$

1.3 Functions

First I want to remind us of some notation and terminology (mathematicians are ***very*** specific).

NOTE To say a function is *on* X *into* Y means that the function is from X to Y.

Definition 1.3.1 (Functions)

We all know what functions are at some level, but here we are going to be extraordinarily explicit. A ***function*** from X into Y is a relation $f \subseteq X \times Y$ such that

(a) $\operatorname{dom} f = X$, and
(b) if $(x, y) \in f$ and $(x, z) \in f$, then $y = z$.

Thus for each $x \in X$ there is exactly one $y \in Y$ with $(x, y) \in f$, and we write $f(x) = y$ (instead of $(x, y) \in f$ or xfy). We call y the ***value*** that f ***assumes*** at the ***argument*** x, and say that f ***maps*** (sends, transforms) x to y, written $x \mapsto y$. We denote the function by

$$f : X \to Y,$$

and call X its ***domain*** and Y its ***codomain***. We use the terms map, mapping, transformation, correspondence, operator, and function interchangeably.

The set of all functions from X to Y is a subset of the power set $\mathcal{P}(X \times Y)$ and is denoted Y^X.

Note that a function is not an active entity, despite action words such as *transforms*; it merely *is*. A function is a static set of pairs, much like a directory that pairs names with addresses.

While $\operatorname{dom} f = X$, the range of f need not be all of Y: $\operatorname{ran} f$ consists of those $y \in Y$ such that $f(x) = y$ for some $x \in X$. More generally, for $A \subseteq X$, the ***image*** of A under f is

$$f(A) = \{f(x) \mid x \in A\} = \{y \in Y \mid f(x) = y \text{ for some } x \in A\},$$

so that $\operatorname{ran} f = f(X)$. We also write $\operatorname{Im} f = f(X)$ and call it the ***image*** of f.

Definition 1.3.2 (Injective, surjective, and bijective maps)

Let $f : X \to Y$.

(a) f is ***surjective***, or maps X ***onto*** Y, if $\operatorname{ran} f = Y$; that is, for every $y \in Y$ there is an $x \in X$ with $f(x) = y$. We denote surjectivity by $f : X \twoheadrightarrow Y$.

(b) f is ***injective***, or ***one-to-one***, if for all $a, b \in X$, $a \neq b$ implies $f(a) \neq f(b)$; equivalently, $f(a) = f(b)$ implies $a = b$. We denote injectivity by $f : X \hookrightarrow Y$.

(c) f is ***bijective***, or a ***one-to-one correspondence***, if it is both injective and surjective. There is no universally agreed notation for bijections; we write $f : X \leftrightarrow Y$ or $f : X \xrightarrow{\sim} Y$.

Definition 1.3.3 (Inclusion and identity maps)

If $X \subseteq Y$, the ***inclusion map*** of X into Y is the function $\iota : X \to Y$ defined by $\iota(x) = x$ for all $x \in X$. The inclusion of X into itself is the ***identity map*** $\mathrm{id}_X : X \to X$, $\mathrm{id}_X(x) = x$; relation-wise, it is equality on X, the set $\{(x, x) \mid x \in X\}$.

Inclusion maps are injective, since $\iota(a) = \iota(b)$ says $a = b$, and identity maps are bijective. An injective map from a set to itself need not be the identity: $n \mapsto n + 1$ is an injection $\mathbb{N} \hookrightarrow \mathbb{N}$, and it is not even surjective.

Definition 1.3.4 (Restrictions and extensions)

Let $f : Y \to Z$ and $X \subseteq Y$. The ***restriction*** of f to X is the function $f|_X : X \to Z$ given by $(f|_X)(x) = f(x)$ for all $x \in X$; conversely, f is called an ***extension*** of $f|_X$ to Y. Note that $\operatorname{ran}(f|_X) = f(X)$.

Definition 1.3.5 (Projections)

The ***projections*** of $X \times Y$ onto its factors are the maps

$$\mathrm{pr}_X : X \times Y \to X,\ (x, y) \mapsto x, \qquad \mathrm{pr}_Y : X \times Y \to Y,\ (x, y) \mapsto y.$$

For a relation $R \subseteq X \times Y$, the domain and range are the images of R under the projections: $\operatorname{dom} R = \mathrm{pr}_X(R)$ and $\operatorname{ran} R = \mathrm{pr}_Y(R)$.

Definition 1.3.6 (Characteristic functions)

Let $2 = \{0, 1\}$. If $A \subseteq X$, the ***characteristic function*** of A is the function $\chi_A : X \to 2$ given by

$$\chi_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \in X - A. \end{cases}$$

Proposition 1.3.7 (Subsets and characteristic functions)

The map $\mathcal{P}(X) \to 2^X$, $A \mapsto \chi_A$, is a bijection. Every function $\varphi : X \to 2$ is χ_A for exactly one $A \subseteq X$, namely $A = \{x \in X \mid \varphi(x) = 1\}$.

Proof. A subset A is recovered from its characteristic function as $A = \{x \in X \mid \chi_A(x) = 1\}$; so $\chi_A = \chi_B$ implies $A = B$, and the map is injective. Given $\varphi : X \to 2$, let $A = \{x \in X \mid \varphi(x) = 1\}$. For $x \in A$ we have $\varphi(x) = 1 = \chi_A(x)$, and for $x \notin A$ we have $\varphi(x) = 0 = \chi_A(x)$, because φ takes only the values 0 and 1. So $\varphi = \chi_A$, and the map is surjective. □

Question 1.3.8

Inclusion maps are always injective; projections, except in a few trivial special cases, are not. What are these special cases? That is, when is the projection $\mathrm{pr}_X : X \times Y \to X$ injective?

Solution. The projection pr_X is injective if and only if $X = \emptyset$ or Y has at most one element.
If $X = \emptyset$ or $Y = \emptyset$, then $X \times Y = \emptyset$, and a function with empty domain is injective vacuously. If $Y = \{y_0\}$, then $\mathrm{pr}_X(x, y_0) = \mathrm{pr}_X(x', y_0)$ means $x = x'$, so $(x, y_0) = (x', y_0)$. Conversely, suppose $X \neq \emptyset$ and Y contains two elements $y \neq y'$. For any $x \in X$, the pairs (x, y) and (x, y') are distinct by Proposition 1.1.1, yet both map to x; so pr_X is not injective.

Question 1.3.9

Prove:

(a) $Y^\emptyset$ has exactly one element, namely $\emptyset$, whether or not Y is empty.

(b) If X is nonempty, then $\emptyset^X$ is empty.

Solution.

(a) A pair $(a, b) \in \emptyset \times Y$ would need $a \in \emptyset$, and there is no such a; so $\emptyset \times Y = \emptyset$. Hence $Y^{\emptyset} \subseteq \mathcal{P}(\emptyset \times Y) = \mathcal{P}(\emptyset) = \{\emptyset\}$, so $Y^{\emptyset}$ has at most one element, and the only candidate is $\emptyset$. It remains to check that $\emptyset$ is a function from $\emptyset$ to Y. It is a subset of $\emptyset \times Y$; its domain is $\emptyset$, because it contains no pairs; and the uniqueness condition holds vacuously, again because it contains no pairs. So $Y^{\emptyset} = \{\emptyset\}$, and its element is called the ***empty function***. Some intuition comes from counting: if X has m elements and Y has n, a function $X \to Y$ makes an independent choice among n values for each of the m arguments, so there are n^m of them, and $n^0 = 1$.

(b) Since $X \neq \emptyset$, choose $x \in X$. A function $f : X \to \emptyset$ would contain a pair $(x, f(x)) \in X \times \emptyset$. But $X \times \emptyset = \emptyset$, so no such function exists, and $\emptyset^X = \emptyset$.

1.4 Families

The notation for families takes some getting used to, since a function is written like an element and its values like subscripts. Consider a function $x : I \to X$. In this setting we call I the ***index set*** and each $i \in I$ an ***index***; the range of the function is called the ***indexed set***; the function itself is called a ***family***; and the value of x at i is called a ***term*** of the family, denoted x_i. We write $\{x_i\}_{i \in I}$, or simply $\{x_i\}$, for the family, and call it a family in X. In particular, a family $\{A_i\}$ of subsets of X is a function $A : I \to \mathcal{P}(X)$, $i \mapsto A_i$. (Strange as this notation is at first, it soon becomes the preferable one.)

Definition 1.4.1 (Unions and intersections of families)

Let $\{A_i\}_{i \in I}$ be a family of sets. Its ***union*** and, when $I \neq \emptyset$, its ***intersection*** are

$$\bigcup_{i \in I} A_i = \{x \mid x \in A_i \text{ for some } i \in I\}, \qquad \bigcap_{i \in I} A_i = \{x \mid x \in A_i \text{ for all } i \in I\}.$$

When $I = \{1, \dots, n\}$ we write $\bigcup_{i=1}^{n} A_i = A_1 \cup \cdots \cup A_n$, and similarly for intersections, just as for sums and products.

The range of a family $\{A_i\}$ is the collection $\{A_i \mid i \in I\}$, and its union is $\bigcup_{i \in I} A_i$. Every collection $\mathcal{C}$ of sets is the range of some family: take $I = \mathcal{C}$ and the family $\mathcal{C} \to \mathcal{C}$, $A \mapsto A$. So nothing is lost by working with families, and the index set is a convenient bookkeeping device.

Example 1.4.2

For the index set $I = \{1, 2, 3\}$ we have $\bigcup_{i \in I} A_i = A_1 \cup A_2 \cup A_3$. Another example may be given as follows: let $I = \mathbb{N}$ and $A_i = \{-i, 0, i\}$, so that $A_0 = \{0\}$, $A_1 = \{-1, 0, 1\}$, $A_2 = \{-2, 0, 2\}$, and so on. The union of the A_i is $\mathbb{Z}$, and their intersection is $\{0\}$.

Example 1.4.3

Let us now consider an infinite index set in the plane: $I = [-1, 1]$ and $A_i = \{i\} \times [0, 1] = \{(i, y) \mid 0 \leqslant y \leqslant 1\} \subseteq \mathbb{R}^2$. Each A_i is a vertical segment. Geometrically, the union of the A_i is the rectangle $[-1, 1] \times [0, 1]$; their intersection is $\emptyset$, since the segments are parallel and pairwise disjoint.

Theorem 1.4.4 (Associative law of unions)

Let $\{I_j\}_{j \in J}$ be a family of sets with $K = \bigcup_{j \in J} I_j$, and let $\{A_k\}_{k \in K}$ be a family of sets. Then

$$\bigcup_{k \in K} A_k = \bigcup_{j \in J} \left(\bigcup_{k \in I_j} A_k \right).$$

If J and every I_j are nonempty, the same holds with intersections in place of unions.

Proof. An element x lies in the left side if and only if $x \in A_k$ for some $k \in K$, that is, for some $j \in J$ and some $k \in I_j$. This says exactly that $x \in \bigcup_{k \in I_j} A_k$ for some $j \in J$, which is membership in the right side. For intersections, x lies in A_k for every $k \in K$ if and only if, for every $j \in J$, it lies in A_k for every $k \in I_j$. □

Proposition 1.4.5 (De Morgan's laws and distributive laws)

Let $\{A_i\}_{i \in I}$ and $\{B_j\}_{j \in J}$ be families of subsets of a set X, with I and J nonempty. Then

$$X - \bigcup_{i \in I} A_i = \bigcap_{i \in I} (X - A_i), \qquad X - \bigcap_{i \in I} A_i = \bigcup_{i \in I} (X - A_i),$$

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

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

Proof. An element of X lies in no A_i if and only if it lies in every $X - A_i$, and it fails to lie in every A_i if and only if it lies in some $X - A_i$; these are De Morgan's laws. For the first distributive law, x lies in the left side if and only if $x \in A_i$ for some i and $x \in B_j$ for some j, that is, $x \in A_i \cap B_j$ for some pair (i, j). For the second, if x lies in every A_i or in every B_j, then $x \in A_i \cup B_j$ for every pair (i, j). Conversely, suppose $x \in A_i \cup B_j$ for every pair and x is not in every A_i, say $x \notin A_{i_0}$. Then $x \in A_{i_0} \cup B_j$ forces $x \in B_j$ for every j. □

Families also let us form products of more than two sets.

Definition 1.4.6 (Products of families)

Let $\{X_i\}_{i \in I}$ be a family of sets. Their ***(Cartesian) product*** $\prod_{i \in I} X_i$ is the set of all families $\{x_i\}_{i \in I}$ with $x_i \in X_i$ for every $i \in I$; that is, of all functions $x : I \to \bigcup_{i \in I} X_i$ with $x(i) \in X_i$ for each i. We also write its elements as $(x_i)_{i \in I}$. When every X_i is the same set X, the product is X^I, the set of all functions $I \to X$.

This generalizes the Cartesian product of two sets. Let $I = \{a, b\}$ with $a \neq b$, and let $X_a = X$ and $X_b = Y$. The map $\prod_{i \in I} X_i \to X \times Y$, $z \mapsto (z_a, z_b)$, is a bijection, since a family indexed by I is determined by its two terms and any choice of the two terms defines one. In the same way, an n-tuple $(x_1, \ldots, x_n)$ is a family indexed by $\{1, \ldots, n\}$, and $X^n = X^{\{1,\ldots,n\}}$; an infinite sequence $(x_0, x_1, x_2, \ldots)$ in X is a family indexed by $\mathbb{N}$, an element of $X^{\mathbb{N}}$. There is more one can say about families of sets, sets of families, and their products, but we will not need it.

1.5 Inverses and Composites

A function $f : X \to Y$ induces a map on subsets, $\mathcal{P}(X) \to \mathcal{P}(Y)$, $A \mapsto f(A)$: just as each element of X is sent to an element of Y, each subset of X is sent to its image. On its own this map is mostly a bookkeeping device, but it sets up its counterpart in the opposite direction, which we use constantly.

Definition 1.5.1 (Inverse images)

Let $f : X \to Y$. For $B \subseteq Y$, the ***inverse image*** of B under f is

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

This defines a map $f^{-1} : \mathcal{P}(Y) \to \mathcal{P}(X)$. For $y \in Y$, the set $f^{-1}(\{y\}) = \{x \in X \mid f(x) = y\}$, also written $f^{-1}(y)$, is the ***fiber*** of f over y.

Surjectivity and injectivity can be read off from inverse images: f is surjective if and only if the inverse image of every nonempty subset of Y is nonempty (equivalently, every fiber is nonempty), and f is injective if and only if the inverse image of each singleton in the range of f is a singleton in X (equivalently, every fiber has at most one element).

There are some connections between images and inverse images which we collect here.

Proposition 1.5.2 (Images and inverse images)

Let $f : X \to Y$.

(a) If $B \subseteq Y$, then $f(f^{-1}(B)) \subseteq B$.

(b) If f is surjective and $B \subseteq Y$, then $f(f^{-1}(B)) = B$.

(c) If $A \subseteq X$, then $A \subseteq f^{-1}(f(A))$.

(d) If f is injective and $A \subseteq X$, then $f^{-1}(f(A)) = A$.

(e) If $\{B_i\}_{i\in I}$ is a family of subsets of Y, then $f^{-1}(\bigcup_{i\in I} B_i) = \bigcup_{i\in I} f^{-1}(B_i)$ and, for $I \neq \emptyset$, $f^{-1}(\bigcap_{i\in I} B_i) = \bigcap_{i\in I} f^{-1}(B_i)$.

(f) $f^{-1}(Y - B) = X - f^{-1}(B)$ for each $B \subseteq Y$.

Proof. (a) An element of $f(f^{-1}(B))$ has the form $f(x)$ with $f(x) \in B$.

(b) If $y \in B$, choose x with $f(x) = y$; then $x \in f^{-1}(B)$, so $y \in f(f^{-1}(B))$. Together with (a), this gives equality.

(c) If $x \in A$, then $f(x) \in f(A)$, so $x \in f^{-1}(f(A))$.

(d) If $x \in f^{-1}(f(A))$, then $f(x) = f(a)$ for some $a \in A$, so $x = a \in A$. Together with (c), this gives equality.

(e) $f(x)$ lies in some B_i if and only if x lies in some $f^{-1}(B_i)$, and $f(x)$ lies in every B_i if and only if x lies in every $f^{-1}(B_i)$.

(f) $f(x) \notin B$ if and only if $x \notin f^{-1}(B)$. □

We now move on to ***composites***.

Definition 1.5.3 (Composition)

Let $f : X \to Y$ and $g : Y \to Z$. Since the codomain of f is the domain of g, we can apply g to the values of f. The ***composite*** of f and g is the function

$$g \circ f : X \to Z, \qquad (g \circ f)(x) = g(f(x)),$$

often written simply gf. We read it right to left: $g \circ f$ applies f first, then g.

Proposition 1.5.4 (Associativity of composition)
Let $f : X \to Y$, $g : Y \to Z$, and $h : Z \to W$. Then $h \circ (g \circ f) = (h \circ g) \circ f$. Moreover, $f \circ \mathrm{id}_X = f = \mathrm{id}_Y \circ f$.

Proof. Both $h \circ (g \circ f)$ and $(h \circ g) \circ f$ are functions from X to W that send each x to $h(g(f(x)))$, so they are the same set of pairs. Similarly, $f \circ \mathrm{id}_X$ and $\mathrm{id}_Y \circ f$ both send x to $f(x)$. □

Composition need not be commutative, even when both composites are defined and have the same domain and codomain: for $f, g : \mathbb{N} \to \mathbb{N}$ with $f(n) = n + 1$ and $g(n) = 2n$, we have $(g \circ f)(n) = 2n + 2$ but $(f \circ g)(n) = 2n + 1$.

Proposition 1.5.5 (Inverse functions)
Let $f : X \to Y$.

(a) If f is bijective, then for each $y \in Y$ there is exactly one $x \in X$ with $f(x) = y$. Setting $f^{-1}(y) = x$ defines a function $f^{-1} : Y \to X$, the ***inverse*** of f, which satisfies $f^{-1} \circ f = \mathrm{id}_X$ and $f \circ f^{-1} = \mathrm{id}_Y$.

(b) Conversely, if there is a function $g : Y \to X$ with $g \circ f = \mathrm{id}_X$ and $f \circ g = \mathrm{id}_Y$, then f is bijective and $g = f^{-1}$.

In particular, the inverse of a bijection is a bijection, with $(f^{-1})^{-1} = f$.

Proof. (a) Such an x exists because f is surjective, and it is unique because f is injective; so f^{-1} is a function. By construction $f(f^{-1}(y)) = y$, and $f^{-1}(f(x))$ is the unique element sent to $f(x)$, namely x.

(b) If $f(a) = f(b)$, then $a = g(f(a)) = g(f(b)) = b$, so f is injective; and every $y \in Y$ equals $f(g(y))$, so f is surjective. Finally, $g(y)$ is an element that f sends to y, so $g(y) = f^{-1}(y)$.

The last statement follows by applying (b) to f^{-1}, with $g = f$. □

NOTE The notation f^{-1} now has two meanings: the inverse image map $\mathcal{P}(Y) \to \mathcal{P}(X)$, defined for every f, and the inverse function $Y \to X$, defined only when f is bijective. For bijective f the two agree in the sense that $f^{-1}(\{y\}) = \{f^{-1}(y)\}$. If f is only injective, the construction in (a) still gives an inverse function $f^{-1} : \operatorname{ran} f \to X$, defined on the range of f.

Proposition 1.5.6 (Composites of injections, surjections, and bijections)
Let $f : X \to Y$ and $g : Y \to Z$.

(a) If f and g are injective, so is $g \circ f$; if they are surjective, so is $g \circ f$. If they are bijective, so is $g \circ f$, and $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.
(b) For inverse images, $(g \circ f)^{-1}(C) = f^{-1}(g^{-1}(C))$ for every $C \subseteq Z$, for arbitrary f and g.

Proof. (a) If $g(f(a)) = g(f(b))$, injectivity of g and then of f gives $a = b$. If $z \in Z$, choose $y \in Y$ with $g(y) = z$ and $x \in X$ with $f(x) = y$; then $g(f(x)) = z$. If both maps are bijective, associativity gives

$$(f^{-1} \circ g^{-1}) \circ (g \circ f) = f^{-1} \circ (g^{-1} \circ g) \circ f = f^{-1} \circ f = \mathrm{id}_X,$$

and similarly $(g \circ f) \circ (f^{-1} \circ g^{-1}) = \mathrm{id}_Z$; now apply Proposition 1.5.5(b).
(b) An element x lies in either side exactly when $g(f(x)) \in C$. □

In both statements the order reverses: to undo "first f, then g" we must undo g first, and the composite $g^{-1} \circ f^{-1}$ is in general not even defined. We will meet the same reversal for inverses in a group (Proposition 2.1.5).

1.5.1 Converse and Composite Relations

Functions are relations, after all, and both constructions of this section make sense for relations in general. The ***converse*** (or inverse) of a relation $R \subseteq X \times Y$ is the relation

$$R^{-1} = \{(y, x) \in Y \times X \mid xRy\}.$$

A function $f : X \to Y$ is bijective exactly when its converse $f^{-1} \subseteq Y \times X$ is a function from Y to X: the domain of the converse is $\operatorname{ran} f$, which is Y exactly when f is surjective, and the uniqueness condition for the converse is injectivity of f. In that case the converse is the inverse function of Proposition 1.5.5.

Composition generalizes in the same way. If $R \subseteq X \times Y$ and $S \subseteq Y \times Z$ are relations, their ***composite***, also called their ***relative product***, is the relation

$$S \circ R = \{(x, z) \in X \times Z \mid xRy \text{ and } ySz \text{ for some } y \in Y\}.$$

When R and S are functions, $x\,(S \circ R)\,z$ holds exactly when $z = S(R(x))$; so composition of functions is a special case of the relative product.

1.6 Quotient Sets and Fibers

We return to equivalence relations, now with functions at our disposal. The quotient set $X/\sim$ of Definition 1.2.7 comes with a map from X.

Definition 1.6.1 (The quotient map)

Let $\sim$ be an equivalence relation on X. The ***quotient map*** (or canonical map) of $\sim$ is

$$q : X \to X/\!\sim, \qquad q(a) = [a].$$

The quotient map is surjective, since every element of $X/\!\sim$ has the form $[a] = q(a)$. By Lemma 1.2.5, $q(a) = q(b)$ if and only if $a \sim b$; so the fiber of q over a class is the class itself, $q^{-1}(\{[a]\}) = [a]$. In particular, q is injective exactly when every class has a single element, that is, when $\sim$ is equality on X.

Conversely, every function produces an equivalence relation.

Proposition 1.6.2 (Fibers of a map)

Let $f : X \to Y$, and for $a, b \in X$ set $a \sim_f b$ if $f(a) = f(b)$.

(a) $\sim_f$ is an equivalence relation on X, and the class of a is the fiber $f^{-1}(f(a))$. Thus the nonempty fibers $f^{-1}(y)$, $y \in \operatorname{ran} f$, form a partition of X.

(b) The map $\bar{f} : X/\!\sim_f \to \operatorname{ran} f$, $[a] \mapsto f(a)$, is well defined and bijective. Its inverse sends $y \in \operatorname{ran} f$ to the fiber $f^{-1}(y)$. In particular, if f is surjective, then $X/\!\sim_f$ is in bijection with Y.

Proof. (a) Reflexivity, symmetry, and transitivity of $\sim_f$ follow from the same properties of equality in Y. The class of a is $\{x \in X \mid f(x) = f(a)\} = f^{-1}(f(a))$, and the fiber over y is nonempty exactly when $y \in \operatorname{ran} f$. Proposition 1.2.8 shows that the classes form a partition.

(b) The value $f(a)$ depends only on the class $[a]$: if $[a] = [b]$, then $a \sim_f b$, that is, $f(a) = f(b)$. So $\bar{f} = \{([a], f(a)) \mid a \in X\}$ is a function. It is injective, since $f(a) = f(b)$ means $[a] = [b]$, and surjective, since each $y \in \operatorname{ran} f$ is $f(a) = \bar{f}([a])$ for some a. The class sent to $y = f(a)$ is $[a] = f^{-1}(y)$. □

Putting the pieces together, three kinds of data on a set carry the same information.

Theorem 1.6.3 (Relations, partitions, and surjections)

Let X be a set.

(a) Equivalence relations on X correspond to partitions of X: an equivalence relation $\sim$ corresponds to $X/\!\sim$, and a partition $\mathfrak{C}$ to $\sim_{\mathfrak{C}}$.

(b) Every equivalence relation $\sim$ on X is the relation $\sim_q$ determined by a surjection, namely its quotient map $q : X \twoheadrightarrow X/\!\sim$.

(c) Every surjection $f : X \twoheadrightarrow Y$ determines the equivalence relation $\sim_f$, whose classes are the fibers of f, and f is the quotient map $q : X \to X/\sim_f$ followed by the bijection $\overline{f} : X/\sim_f \to Y$; that is, $f = \overline{f} \circ q$.

So an equivalence relation on X, a partition of X, and a surjection from X onto another set (up to a bijection of the target) are three descriptions of the same thing.

Proof. Part (a) is Proposition 1.2.8. For (b), $q(a) = q(b)$ if and only if $a \sim b$, by Lemma 1.2.5; so $\sim_q$ is $\sim$. For (c), Proposition 1.6.2 shows that $\overline{f}$ is a bijection onto $\operatorname{ran} f = Y$, and $(\overline{f} \circ q)(a) = \overline{f}([a]) = f(a)$ for every $a \in X$. □

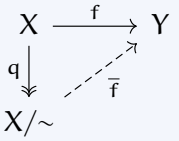
More generally, given an equivalence relation $\sim$ on X and an arbitrary function $f : X \to Y$, we may ask whether f induces a function on the quotient by the rule $[a] \mapsto f(a)$. The difficulty is that a class $[a]$ has many representatives, and the proposed value must not depend on which one we pick.

Proposition 1.6.4 (Maps descending to a quotient)

Let $\sim$ be an equivalence relation on X with quotient map $q : X \to X/\sim$, and let $f : X \to Y$. There is a function $\overline{f} : X/\sim \to Y$ with $f = \overline{f} \circ q$ if and only if f is ***compatible*** with $\sim$, meaning that

$$a \sim b \implies f(a) = f(b) \qquad \text{for all } a, b \in X.$$

In that case $\overline{f}$ is unique, it is given by $\overline{f}([a]) = f(a)$, and its image is $f(X)$.



Proof. If $\overline{f}$ exists and $a \sim b$, then $q(a) = q(b)$, so $f(a) = \overline{f}(q(a)) = \overline{f}(q(b)) = f(b)$. Conversely, suppose f is compatible with $\sim$, and let $\overline{f} = \{([a], f(a)) \mid a \in X\} \subseteq (X/\sim) \times Y$. Its domain is $X/\sim$, and if two of its pairs $([a], f(a))$ and $([b], f(b))$ have the same first coordinate, then $[a] = [b]$, so $a \sim b$ and $f(a) = f(b)$. Hence $\overline{f}$ is a function, and $\overline{f}(q(a)) = \overline{f}([a]) = f(a)$ for all a. It is unique because q is surjective: any g with $f = g \circ q$ satisfies $g([a]) = g(q(a)) = f(a)$ for every class $[a]$. Finally, $\overline{f}(X/\sim) = \{f(a) \mid a \in X\} = f(X)$. □

When f is compatible with $\sim$, we say that f ***descends*** to $X/\sim$, or ***factors through*** q. The diagram above ***commutes***: the two routes from X to Y give the same map. The induced map $\overline{f}$ is injective if and only if $f(a) = f(b)$ implies $a \sim b$, that is, if

and only if $\sim$ is exactly the relation $\sim_f$; and $\bar{f}$ is surjective if and only if f is. This principle will return many times: quotient groups (Theorem 3.3.2) and quotient vector spaces (Definition 7.1.1) are built on it. Here is its first use.

Example 1.6.5 (Addition mod n)

Let $\sim$ be congruence mod n on $\mathbb{Z}$ (Example 1.2.9). We would like to add classes by the rule $[a] + [b] = [a + b]$. For this to make sense, the right side must depend only on the classes $[a]$ and $[b]$, not on the representatives a and b. It does: if $a \sim a'$ and $b \sim b'$, then n divides $a' - a$ and $b' - b$, hence their sum $(a' + b') - (a + b)$, so $a + b \sim a' + b'$. In the language of Proposition 1.6.4: for fixed b, the map $\mathbb{Z} \to \mathbb{Z}/\sim$, $a \mapsto [a + b]$, is compatible with $\sim$, so it descends to a map $\mathbb{Z}/\sim \to \mathbb{Z}/\sim$, $[a] \mapsto [a + b]$; and the result depends only on $[b]$, since $[a + b] = [a + b']$ whenever $b \sim b'$. Hence

$$[a] + [b] \coloneqq [a + b]$$

is a well-defined operation on $\mathbb{Z}/\sim$. Under the bijection $\{0, 1, \ldots, n-1\} \to \mathbb{Z}/\sim$, $r \mapsto [r]$, it becomes addition of remainders: $[a] + [b]$ is the class $[r]$, where r is the remainder of $a + b$ on division by n. This is the group $\mathbb{Z}/n\mathbb{Z}$ of the next chapter (Example 2.2.3), which we describe there as $\{0, 1, \ldots, n - 1\}$ with addition mod n.

1.7 Cardinality

When do two sets have the same size? For finite sets we can count; in general, we compare sets by means of maps between them.

Definition 1.7.1 (Cardinality)

Sets X and Y have the ***same cardinality***, written $|X| = |Y|$, if there exists a bijection $X \to Y$. We write $|X| \leqslant |Y|$ if there exists an injection $X \hookrightarrow Y$, and $|X| < |Y|$ if $|X| \leqslant |Y|$ and $|X| \neq |Y|$.

These are relations between sets; we do not define $|X|$ as an object in its own right, except for finite sets. A set X is ***finite*** if $|X| = |\{1, \ldots, n\}|$ for some $n \in \mathbb{N}$ (where $\{1, \ldots, 0\} = \emptyset$); this n is unique, since an induction on n shows that there is no injection $\{1, \ldots, n + 1\} \hookrightarrow \{1, \ldots, n\}$, and we write $|X| = n$, the number of elements of X. A set X with $|X| = |\mathbb{N}|$ is ***countably infinite***; a set that is finite or countably infinite is ***countable***; any other set is ***uncountable***.

Having the same cardinality behaves like an equivalence relation: id_X is a bijection, the inverse of a bijection is a bijection, and a composite of bijections is a bijection (Propositions 1.5.5 and 1.5.6). Likewise $|X| \leqslant |X|$, and $|X| \leqslant |Y| \leqslant |Z|$ implies $|X| \leqslant |Z|$, since a composite of injections is injective. If $X \subseteq Y$, the inclusion map shows that $|X| \leqslant |Y|$. The notation $\leqslant$ is justified by the following theorem, which lets us relate cardinality and maps between sets.

Theorem 1.7.2 (Schröder–Bernstein)
If there exist injections $f : X \hookrightarrow Y$ and $g : Y \hookrightarrow X$, then $|X| = |Y|$.

Proof. We build a bijection $h : X \to Y$ that uses f on part of X and the inverse of g on the rest. Since g is injective, every $x \in g(Y)$ is $g(y)$ for exactly one $y \in Y$, which we denote $g^{-1}(x)$.

Define subsets $C_0, C_1, C_2, \ldots$ of X by $C_0 = X - g(Y)$ and $C_{k+1} = g(f(C_k))$, and let $C = \bigcup_{k \in \mathbb{N}} C_k$. Every $x \in X - C$ lies outside C_0, so $x \in g(Y)$. Hence

$$h(x) = \begin{cases} f(x) & \text{if } x \in C, \\ g^{-1}(x) & \text{if } x \in X - C \end{cases}$$

defines a function $h : X \to Y$.

The map h is injective. Let $x \neq x'$. If both lie in C, then $h(x) = f(x) \neq f(x') = h(x')$. If neither does, then $h(x) = g^{-1}(x)$ and $h(x') = g^{-1}(x')$ are distinct, since g sends them to x and x'. Suppose $x \in C$ and $x' \notin C$, say $x \in C_k$. If $h(x) = h(x')$, then $f(x) = g^{-1}(x')$, so $x' = g(f(x)) \in g(f(C_k)) = C_{k+1} \subseteq C$, a contradiction.

The map h is surjective. Let $y \in Y$. If $g(y) \notin C$, then $h(g(y)) = g^{-1}(g(y)) = y$. If $g(y) \in C$, then $g(y) \notin C_0$, because $g(y) \in g(Y)$; so $g(y) \in C_{k+1} = g(f(C_k))$ for some k, and there is an $x \in C_k$ with $g(y) = g(f(x))$. Since g is injective, $y = f(x) = h(x)$. □

NOTE The set C has a simple description. Start from a point $x_0 \in X$ that is not in the image of g and follow the chain $x_0 \mapsto f(x_0) \mapsto g(f(x_0)) \mapsto f(g(f(x_0))) \mapsto \cdots$; then C consists of the points of X reached by such chains. On these points h follows f forward, and everywhere else h follows g backward.

Several sets that look larger than $\mathbb{N}$ turn out to have the same cardinality.

Proposition 1.7.3 (The integers are countable)

The map $f : \mathbb{Z} \to \mathbb{N}$ given by

$$f(n) = \begin{cases} 2n & \text{if } n \geqslant 0, \\ -(2n+1) & \text{if } n < 0 \end{cases}$$

is a bijection. Hence $|\mathbb{Z}| = |\mathbb{N}|$.

Proof. The map f sends the nonnegative integers to the even natural numbers and the negative integers to the odd ones ($-1 \mapsto 1$, $-2 \mapsto 3$, and so on). The map $g : \mathbb{N} \to \mathbb{Z}$ given by $g(m) = m/2$ for even m and $g(m) = -(m+1)/2$ for odd m satisfies $g \circ f = \mathrm{id}_{\mathbb{Z}}$ and $f \circ g = \mathrm{id}_{\mathbb{N}}$, so f is a bijection by Proposition 1.5.5. □

Proposition 1.7.4 (Pairs of natural numbers)

The map $\mathbb{N} \times \mathbb{N} \to \mathbb{N}$, $(k, l) \mapsto 2^k(2l+1) - 1$, is a bijection. Consequently $|\mathbb{N}^n| = |\mathbb{N}|$ for every $n \geqslant 1$.

Proof. Every positive integer m can be written as $m = 2^k(2l+1)$ with $k, l \in \mathbb{N}$ in exactly one way: 2^k must be the largest power of 2 dividing m, and $2l+1$ the odd number $m/2^k$. So $(k, l) \mapsto 2^k(2l+1)$ is a bijection from $\mathbb{N} \times \mathbb{N}$ onto the positive integers, and the additional -1 shifts the positive integers onto $\mathbb{N}$; for instance, $(0, 0) \mapsto 2^0(2 \cdot 0 + 1) - 1 = 0$.

For the second statement we induct on n, the case $n = 1$ being trivial. Sending $(x_1, \ldots, x_{n+1})$ to $((x_1, \ldots, x_n), x_{n+1})$ is a bijection $\mathbb{N}^{n+1} \to \mathbb{N}^n \times \mathbb{N}$. If $\beta : \mathbb{N}^n \to \mathbb{N}$ is a bijection, then $(u, v) \mapsto (\beta(u), v)$ is a bijection $\mathbb{N}^n \times \mathbb{N} \to \mathbb{N} \times \mathbb{N}$, with inverse $(a, v) \mapsto (\beta^{-1}(a), v)$, and the first part finishes the induction. □

Proposition 1.7.5 (The rationals are countable)

$|\mathbb{Q}| = |\mathbb{N}|$. More generally, $\mathbb{Z}^n$ and $\mathbb{Q}^n$ are countably infinite for every $n \geqslant 1$.

Proof. For $x \in \mathbb{Q}$, let b be the smallest positive integer such that $bx \in \mathbb{Z}$, and send x to the pair $(bx, b) \in \mathbb{Z} \times \mathbb{N}$. This is injective, since x is recovered as bx/b. Composing with the bijection $\mathbb{Z} \times \mathbb{N} \to \mathbb{N} \times \mathbb{N}$, $(a, b) \mapsto (f(a), b)$, where f is the bijection of Proposition 1.7.3, and then with the bijection of Proposition 1.7.4, gives an injection $\mathbb{Q} \hookrightarrow \mathbb{N}$. The inclusion $\mathbb{N} \subseteq \mathbb{Q}$ gives an injection the other way, and the Schröder–Bernstein theorem gives $|\mathbb{Q}| = |\mathbb{N}|$.

If $\beta : \mathbb{Q} \to \mathbb{N}$ is a bijection, then $(x_1, \ldots, x_n) \mapsto (\beta(x_1), \ldots, \beta(x_n))$ is a bijection $\mathbb{Q}^n \to \mathbb{N}^n$, and $|\mathbb{N}^n| = |\mathbb{N}|$ by Proposition 1.7.4. The same argument applies to $\mathbb{Z}^n$. □

Not every infinite set is countable. To show this for $\mathbb{R}$ we use decimal expansions, and one fact about them from analysis: every $t \in [0,1)$ has exactly one decimal expansion

$$t = 0.a_0a_1a_2\ldots = \sum_{k\geqslant 0} a_k 10^{-(k+1)}, \qquad a_k \in \{0,1,\ldots,9\},$$

that does not end in an infinite string of 9s (we exclude, for instance, $0.0999\ldots$, which equals $0.1000\ldots$); and every digit sequence that does not end in an infinite string of 9s is the expansion of some $t \in [0,1)$. We call this expansion the ***decimal expansion*** of t.

Theorem 1.7.6 (Cantor's diagonal argument)
No map $\mathbb{N} \to \mathbb{R}$ is surjective. Hence $|\mathbb{N}| < |\mathbb{R}|$, and $\mathbb{R}$ is uncountable.

Proof. We first show that no map $g : \mathbb{N} \to [0,1)$ is surjective. Write the decimal expansions of its values one below the other,

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

and go down the diagonal: let $b_j = 5$ if $a_{jj} \neq 5$ and $b_j = 4$ if $a_{jj} = 5$, and let $y = 0.b_0b_1b_2\ldots$. All digits of y are 4s and 5s, so $y \in [0,1)$ and this is its decimal expansion. For each j, the decimal expansions of y and $g(j)$ differ in the jth digit, so $y \neq g(j)$ by uniqueness of decimal expansions. Thus y is not in the range of g.

Now let $f : \mathbb{N} \to \mathbb{R}$, and define $g : \mathbb{N} \to [0,1)$ by $g(j) = f(j)$ if $f(j) \in [0,1)$ and $g(j) = 0$ otherwise. Every element of $[0,1)$ in the range of f is in the range of g. So the element $y \in [0,1)$ just constructed for g, which is not in the range of g, is not in the range of f either, and f is not surjective. In particular there is no bijection $\mathbb{N} \to \mathbb{R}$, while the inclusion $\mathbb{N} \subseteq \mathbb{R}$ gives $|\mathbb{N}| \leqslant |\mathbb{R}|$. □

The same idea shows that there are arbitrarily large cardinalities.

Theorem 1.7.7 (There is no largest cardinality)
For every set S, no map $S \to \mathcal{P}(S)$ is surjective. Hence $|S| < |\mathcal{P}(S)|$.

Proof. Let $\alpha : S \to \mathcal{P}(S)$, and let $A = \{s \in S \mid s \notin \alpha(s)\}$. Suppose $A = \alpha(t)$ for some $t \in S$. Then $t \in A$ if and only if $t \notin \alpha(t) = A$, a contradiction. So A is not in

the range of α, and there is no bijection $S \to \mathcal{P}(S)$. On the other hand, $s \mapsto \{s\}$ is an injection $S \hookrightarrow \mathcal{P}(S)$. □

This is the diagonal argument again. Think of a map $\alpha : S \to \mathcal{P}(S)$ as a table with rows and columns indexed by S, whose entry in row s and column t is 1 if $t \in \alpha(s)$ and 0 otherwise. The set A is obtained by going down the diagonal and switching every entry, so it differs from the row of s in column s, for every s.

By Proposition 1.3.7, subsets of S and functions $S \to \{0,1\}$ amount to the same thing: $A \mapsto \chi_A$ is a bijection $\mathcal{P}(S) \to \{0,1\}^S$, whose inverse sends φ to $\varphi^{-1}(1)$. So $|\{0,1\}^S| = |\mathcal{P}(S)| > |S|$. If S is finite with $|S| = n$, a function $S \to \{0,1\}$ makes one of two choices at each of the n elements, so $|\mathcal{P}(S)| = 2^n$, and the theorem generalizes the inequality $n < 2^n$. Applying the theorem repeatedly produces ever larger sets S, $\mathcal{P}(S)$, $\mathcal{P}(\mathcal{P}(S))$, ...; so there is no largest cardinality.

Finally, sequences of real numbers are no more numerous than real numbers themselves.

Proposition 1.7.8 (Cardinality of mappings)
$|\mathbb{R}^{\mathbb{N}}| = |\mathbb{R}|$.

Proof. We find injections in both directions and apply the Schröder–Bernstein theorem, working with $[0,1)$ in place of $\mathbb{R}$.

First, $|\mathbb{R}| = |[0,1)|$. The inclusion $[0,1) \subseteq \mathbb{R}$ is an injection one way. For the other, $x \mapsto x/(1+|x|)$ is a strictly increasing, hence injective, map $\mathbb{R} \to (-1,1)$, and $u \mapsto (u+1)/2$ maps $(-1,1)$ injectively into $(0,1) \subseteq [0,1)$. By the Schröder–Bernstein theorem, there is a bijection $\beta : \mathbb{R} \to [0,1)$. Applying β termwise, $(x_j)_{j\in\mathbb{N}} \mapsto (\beta(x_j))_{j\in\mathbb{N}}$, is a bijection $\mathbb{R}^{\mathbb{N}} \to [0,1)^{\mathbb{N}}$, with inverse given by applying β^{-1} termwise. So it suffices to show that $|[0,1)^{\mathbb{N}}| = |[0,1)|$.

The map $t \mapsto (t,0,0,\ldots)$ is an injection $[0,1) \hookrightarrow [0,1)^{\mathbb{N}}$. For the other direction we interleave decimal digits. Let $\pi : \mathbb{N}\times\mathbb{N} \to \mathbb{N}$, $\pi(j,k) = 2^j(2k+1) - 1$, be the bijection of Proposition 1.7.4. Given a sequence $(t_j)_{j\in\mathbb{N}}$ in $[0,1)$, write the decimal expansions $t_j = 0.a_{j0}a_{j1}a_{j2}\ldots$, and define a digit sequence $(c_m)_{m\in\mathbb{N}}$ by

$$c_{\pi(j,k)} = a_{jk};$$

that is, the kth digit of t_j is placed in position $\pi(j,k)$. Since π is a bijection, every position receives exactly one digit. The positions $\pi(0,k) = 2k$ carry the digits of t_0, infinitely many of which differ from 9; so (c_m) does not end in an infinite string of 9s, and it is the decimal expansion of some $y \in [0,1)$. If $(t_j) \neq (t'_j)$, then $t_j \neq t'_j$ for some j, so their decimal expansions differ in some digit k, and the two digit sequences built from them differ in position $\pi(j,k)$; by uniqueness of decimal expansions, the corresponding values of y differ. So $(t_j) \mapsto y$ is an injection $[0,1)^{\mathbb{N}} \hookrightarrow [0,1)$.

By the Schröder–Bernstein theorem, $|[0,1)^{\mathbb{N}}| = |[0,1)|$, and therefore $|\mathbb{R}^{\mathbb{N}}| = |[0,1)^{\mathbb{N}}| = |[0,1)| = |\mathbb{R}|$. □

CHAPTER 2

Groups and Homomorphisms

Groups are the central object of abstract algebra, capturing the essence of symmetry and invertible operations. The definition is deceptively simple—three axioms—yet groups appear everywhere: in geometry (symmetries), number theory (modular arithmetic), topology (fundamental groups), physics (gauge theories), and even chemistry (molecular symmetries). This chapter lays the foundations: the axioms and their first consequences, the fundamental examples, subgroups, homomorphisms, and the order of an element.

2.1 The Group Axioms

Definition 2.1.1 (Groups)

A ***group*** is a set G together with a map

$$m : G \times G \to G$$

called a ***law of composition*** (also ***binary operation***). Writing ab for $m(a, b)$, the law of composition must satisfy three axioms:

(a) **Associativity:** $\forall a, b, c \in G$, $(ab)c = a(bc)$.

(b) **Identity:** $\exists e \in G$ such that $ea = ae = a$ for all $a \in G$.

(c) **Inverses:** $\forall a \in G$, $\exists b \in G$ such that $ab = ba = e$.

We write (G, m) or simply G for the group, and call its cardinality $|G|$ (Section 1.7) the ***order*** of G. The group is ***finite*** or ***infinite*** according as the set G is.

The axioms have immediate but important consequences, which we now prove carefully.

Proposition 2.1.2 (Uniqueness of identity)

The identity element e is unique.

Proof. Suppose e and e' are both identities. Then $e = ee'$ (since e' is an identity) and $ee' = e'$ (since e is an identity). Thus $e = e'$. □

Proposition 2.1.3 (Uniqueness of inverses)

For each $a \in G$, the inverse is unique. We denote it a^{-1} (or $-a$ in additive notation).

Proof. Suppose b and c are both inverses of a, so $ab = ba = e$ and $ac = ca = e$. Then

$$b = be = b(ac) = (ba)c = ec = c.$$

□

Proposition 2.1.4 (Cancellation laws)

In a group G:

(a) **Left cancellation:** $ab = ac \implies b = c$.

(b) **Right cancellation:** $ba = ca \implies b = c$.

Proof. For left cancellation: if $ab = ac$, multiply both sides on the left by a^{-1}: $a^{-1}(ab) = a^{-1}(ac)$, so $(a^{-1}a)b = (a^{-1}a)c$, giving $eb = ec$, i.e., $b = c$. Right cancellation is similar, multiplying on the right by a^{-1}. □

Proposition 2.1.5 (Inverse of products)

For all $a, b \in G$: $(ab)^{-1} = b^{-1}a^{-1}$. Note the reversal of order!

Proof. We verify $(ab)(b^{-1}a^{-1}) = e$: $(ab)(b^{-1}a^{-1}) = a(bb^{-1})a^{-1} = aea^{-1} = aa^{-1} = e$. Similarly, $(b^{-1}a^{-1})(ab) = b^{-1}(a^{-1}a)b = b^{-1}b = e$. By uniqueness of inverses, $(ab)^{-1} = b^{-1}a^{-1}$. □

Proposition 2.1.6 (Inverse of inverse)

For all $a \in G$: $(a^{-1})^{-1} = a$.

Proof. By definition, a^{-1} satisfies $aa^{-1} = a^{-1}a = e$. This says that a is an inverse of a^{-1}, so by uniqueness, $(a^{-1})^{-1} = a$. □

By associativity, a product $a_1 a_2 \cdots a_k$ of any number of elements has the same value however we insert parentheses (an induction on k proves this), so we omit them. In particular, powers of an element make sense.

Definition 2.1.7 (Powers)

For $a \in G$ and $n \in \mathbb{Z}$, set $a^0 = e$; for $n > 0$, let $a^n = a^{n-1}a$, the product of n copies of a; and for $n < 0$, let $a^n = (a^{-1})^{-n}$.

The usual laws of exponents hold: $a^m a^n = a^{m+n}$ and $(a^m)^n = a^{mn}$ for all $m, n \in \mathbb{Z}$, and in particular $(a^n)^{-1} = a^{-n}$. For $m, n \geqslant 0$ these follow by induction on n, and the remaining cases reduce to these by writing negative powers of a as positive powers of a^{-1}; we omit the routine verification. If $ab = ba$, then also $(ab)^n = a^n b^n$.

2.1.1 Group Variants and Related Structures

Groups sit in a hierarchy of algebraic structures with fewer or more axioms.

Definition 2.1.8 (Abelian groups)

A group G is ***abelian*** (or ***commutative***) if $ab = ba$ for all $a, b \in G$.

NOTE It is customary to use additive notation ($+$, 0, $-a$, and na for the nth power) for abelian groups, and multiplicative notation ($\cdot$, 1 or e, a^{-1}) in general. Multiplicative notation does *not* imply that the group is non-abelian!

Definition 2.1.9 (Monoids)

A ***monoid*** is a set with an associative binary operation and an identity element; inverses are not required. Examples: $(\mathbb{N}, +)$ with identity 0, and $(\mathbb{Z}, \times)$ with identity 1.

Definition 2.1.10 (Semigroups)

A ***semigroup*** is a set with an associative binary operation; neither an identity nor inverses are required. Example: $(\mathbb{Z}_{>0}, +)$.

NOTE There is a pattern here: group = monoid + inverses = semigroup + identity + inverses. Category theory (Definition 8.1.1) provides a unifying framework: a monoid is a category with one object, and a group is a category with one object in which every morphism is invertible.

2.2 Fundamental Examples

Groups, like sets, can be finite, countably infinite, or uncountable (Section 1.7); our examples include all three kinds.

Example 2.2.1 (The trivial group)

The one-element set $\{e\}$ with $ee = e$ is a group, the ***trivial group***. It is rarely an interesting example.

2.2.1 Integers and Modular Arithmetic

Example 2.2.2 (The integers)

The set $(\mathbb{Z}, +)$ is an abelian group with identity 0 and inverse $-n$ for each n; so are $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ under addition. The natural numbers $(\mathbb{N}, +)$, by contrast, form only a monoid, since 1 has no inverse in $\mathbb{N}$. We will see that $\mathbb{Z}$ is the prototypical infinite cyclic group (Theorem 2.6.2).

Example 2.2.3 (Integers mod n)

For an integer $n \geqslant 1$, let $\mathbb{Z}/n\mathbb{Z} = \{0, 1, 2, \ldots, n-1\}$, also written $\mathbb{Z}/n$, with ***addition mod*** n:

$$a +_n b \coloneqq \begin{cases} a + b & \text{if } a + b < n, \\ a + b - n & \text{if } a + b \geqslant n. \end{cases}$$

Thus $a +_n b$ is the remainder of $a + b$ on division by n. The identity is 0, and the inverse of $a \neq 0$ is $n - a$. For associativity, note that $a +_n b$ differs from $a + b$ by a multiple of n; hence $(a +_n b) +_n c$ and $a +_n (b +_n c)$ are both remainders of $a + b + c$ on division by n, and remainders are unique (Example 1.2.9). So $\mathbb{Z}/n\mathbb{Z}$ is an abelian group of order n. We usually write $+$ for $+_n$.

In Example 1.6.5 we met the same group as the quotient set of $\mathbb{Z}$ by congruence mod n, with $[a] + [b] = [a + b]$; the bijection $r \mapsto [r]$ matches the two descriptions.

The quotient description is the better one, and the next chapter develops it for arbitrary groups.

Example 2.2.4 (The circle $\mathbb{R}/\mathbb{Z}$)
Similarly, let $\mathbb{R}/\mathbb{Z} = [0,1) \subseteq \mathbb{R}$ with ***addition mod*** 1:

$$a +_1 b := \begin{cases} a+b & \text{if } a+b < 1, \\ a+b-1 & \text{if } a+b \geqslant 1. \end{cases}$$

Here $a +_1 b$ is the fractional part of $a+b$, and the argument used for $\mathbb{Z}/n\mathbb{Z}$ shows that $\mathbb{R}/\mathbb{Z}$ is an abelian group with identity 0, the inverse of $a \neq 0$ being $1-a$. Wrapping $[0,1)$ around a circle of circumference 1 turns addition mod 1 into the addition of angles (Example 2.5.7). The group $\mathbb{R}/\mathbb{Z}$ reappears in the next chapter as the quotient of $\mathbb{R}$ by its subgroup $\mathbb{Z}$.

2.2.2 Multiplicative Groups

Example 2.2.5 (Nonzero numbers)
$\mathbb{Q}^* = \mathbb{Q} \setminus \{0\}$, $\mathbb{R}^* = \mathbb{R} \setminus \{0\}$, and $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$ are abelian groups under multiplication, with identity 1 and inverse $1/x$.

Example 2.2.6 (The circle group)
The unit circle $S^1 = \{z \in \mathbb{C} : |z| = 1\}$ is a group under multiplication, since $|zw| = |z||w|$ and $|1/z| = 1/|z|$. It is a *compact* abelian group and plays a role in Fourier analysis and representation theory. We will see in Example 2.5.7 that $t \mapsto e^{2\pi i t}$ identifies $\mathbb{R}/\mathbb{Z}$ with S^1.

Example 2.2.7 (Roots of unity)
For each $n \geqslant 1$, the nth roots of unity $\mu_n = \{z \in \mathbb{C} : z^n = 1\} = \{e^{2\pi i k/n} : k = 0, 1, \ldots, n-1\}$ form a group of order n under multiplication: if $z^n = w^n = 1$, then $(zw)^n = 1$ and $(z^{-1})^n = 1$. We will see that μ_n is cyclic and isomorphic to $\mathbb{Z}/n\mathbb{Z}$ (Example 2.6.3).

2.2.3 Symmetric Groups and Cycle Notation

Definition 2.2.8 (Permutation groups)

A ***permutation*** of a set X is a bijection $\sigma : X \to X$. The set of all permutations of X, with composition as the law of composition, is denoted $\mathrm{Sym}(X)$ (also $\mathrm{Perm}(X)$).

Proposition 2.2.9 (Permutations form a group)

$\mathrm{Sym}(X)$ is a group under composition, with identity id_X; the inverse of σ is the inverse function σ^{-1}.

Proof. A composite of bijections is a bijection (Proposition 1.5.6), so composition is a law of composition on $\mathrm{Sym}(X)$, and it is associative by Proposition 1.5.4. The identity map id_X is a bijection with $\sigma \circ \mathrm{id}_X = \sigma = \mathrm{id}_X \circ \sigma$. Each σ has an inverse function σ^{-1}, which is a bijection with $\sigma \circ \sigma^{-1} = \mathrm{id}_X = \sigma^{-1} \circ \sigma$ (Proposition 1.5.5). □

As with all maps, we compose permutations right to left: $\sigma\tau = \sigma \circ \tau$ means apply τ first, then σ.

Definition 2.2.10 (Symmetric group)

The ***symmetric group*** on n letters is $S_n = \mathrm{Sym}(\{1, 2, \ldots, n\})$.

A permutation $\sigma \in S_n$ is determined by the list $\sigma(1), \ldots, \sigma(n)$, which contains each of $1, \ldots, n$ exactly once. There are n choices for $\sigma(1)$, then $n-1$ for $\sigma(2)$, and so on, so $|S_n| = n!$. To write permutations down efficiently we use cycles.

Definition 2.2.11 (Cycles)

Let $a_1, \ldots, a_k$ be distinct elements of $\{1, \ldots, n\}$. The ***cycle*** $(a_1\ a_2\ \cdots\ a_k) \in S_n$ is the permutation that sends

$$a_1 \mapsto a_2 \mapsto \cdots \mapsto a_k \mapsto a_1$$

and fixes every other element. A cycle with k entries is a k***-cycle***, and a 2-cycle $(a\ b)$, which swaps a and b, is a ***transposition***. Two cycles are ***disjoint*** if they have no entries in common.

The same cycle can be written starting from any of its entries: $(1\,2\,3) = (2\,3\,1) = (3\,1\,2)$. Its inverse runs backwards, $(a_1\,a_2\,\cdots\,a_k)^{-1} = (a_k\,\cdots\,a_2\,a_1)$, and a 1-cycle (a) is the identity, usually written e. Disjoint cycles move disjoint sets of elements, so they commute. Every permutation is a product of disjoint cycles; this is proved in the next chapter (Theorem 3.4.1).

Example 2.2.12 (Composing cycles)

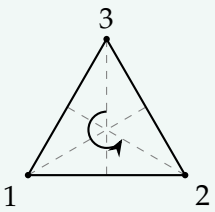
In S_3, compute $(1\,2)(2\,3)$, remembering that the right-hand factor acts first:

$$1 \xmapsto{(2\,3)} 1 \xmapsto{(1\,2)} 2, \qquad 2 \xmapsto{(2\,3)} 3 \xmapsto{(1\,2)} 3, \qquad 3 \xmapsto{(2\,3)} 2 \xmapsto{(1\,2)} 1.$$

So $(1\,2)(2\,3) = (1\,2\,3)$. In the other order, $1 \mapsto 2 \mapsto 3$, $2 \mapsto 1 \mapsto 1$, and $3 \mapsto 3 \mapsto 2$, so $(2\,3)(1\,2) = (1\,3\,2)$. Since the two products differ, S_3 is non-abelian. The same computation in S_n, where the letters other than $1, 2, 3$ are fixed, shows that S_n is non-abelian for every $n \geqslant 3$.

Example 2.2.13 (Symmetries of a triangle)

S_3 can be visualized as the symmetries of an equilateral triangle: the rigid motions of the plane that carry the triangle onto itself. There are 3 rotations (including the identity) and 3 reflections.



Labeling the vertices $1, 2, 3$, each symmetry permutes the vertices: the rotations give e, $(1\,2\,3)$, and $(1\,3\,2)$, and the reflections give $(1\,2)$, $(1\,3)$, and $(2\,3)$. These are six distinct permutations, and $|S_3| = 3! = 6$, so every permutation of the vertices comes from exactly one symmetry. Moreover, composing two symmetries composes the corresponding permutations. So S_3 also occurs as the group of symmetries of the triangle.

2.2.4 Dihedral Groups

The triangle is the first of a family of examples, one for each regular polygon.

Definition 2.2.14 (Dihedral groups)

Let $n \geqslant 3$. The ***dihedral group*** D_n is the group of symmetries of a regular n-gon in the plane: its elements are the rigid motions of the plane that carry the n-gon onto itself, and its law of composition is composition of maps. We write r for the counterclockwise rotation by $2\pi/n$ about the center, and s for the reflection across a fixed line through the center and a vertex. As Proposition 2.2.15 shows, D_n has exactly $2n$ elements, the rotations r^k and the reflections $r^k s$ ($0 \leqslant k \leqslant n-1$), and

$$r^n = e, \qquad s^2 = e, \qquad srs^{-1} = r^{-1}.$$

Composites and inverses of symmetries are symmetries, and composition of maps is associative, so D_n is indeed a group.

Proposition 2.2.15 (Structure of D_n)

The group D_n has exactly $2n$ elements,

$$D_n = \{e, r, r^2, \ldots, r^{n-1}, s, sr, sr^2, \ldots, sr^{n-1}\},$$

where r^k is the rotation by $2\pi k/n$ and the sr^k are the n reflections. These elements satisfy

$$r^n = e, \qquad s^2 = e, \qquad srs^{-1} = r^{-1} \quad (\text{equivalently, } sr = r^{-1}s).$$

Consequently $sr^k = r^{-k}s$ for all k, and the reflections can equally be written $r^k s$ ($0 \leqslant k \leqslant n-1$): $D_n = \{r^k, r^k s \mid 0 \leqslant k \leqslant n-1\}$.

Proof. We use two facts from plane geometry: a symmetry of the n-gon permutes its vertices and carries adjacent vertices to adjacent vertices, and a symmetry is determined by what it does to the vertices. Number the vertices $0, 1, \ldots, n-1$ counterclockwise, with vertex 0 on the axis of s, and read vertex numbers mod n. Then, on vertices,

$$r^m(j) = j + m, \qquad s(j) = -j, \qquad sr^m(j) = -j - m.$$

The first formula gives $r^n = e$, and the second gives $s^2 = e$, so $s^{-1} = s$. Also $srs^{-1}(j) = s(r(s(j))) = s(1-j) = j-1$, so $srs^{-1} = r^{-1}$; multiplying on the right by s gives $sr = r^{-1}s$. Raising $srs^{-1} = r^{-1}$ to the kth power gives $sr^k s^{-1} = r^{-k}$, that is, $sr^k = r^{-k}s$; as k runs over $0, \ldots, n-1$, so does the exponent of $r^{-k} = r^{n-k}$ taken mod n, so $\{sr^k\} = \{r^k s\}$.

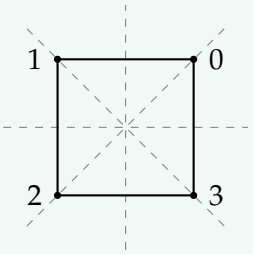
The $2n$ maps $j \mapsto j + m$ and $j \mapsto -j - m$ ($0 \leqslant m \leqslant n-1$) are distinct: two maps of the same kind differ at $j = 0$, and if $j + m \equiv -j - m'$ for all j, then taking

$j = 0$ and $j = 1$ gives $2 \equiv 0 \pmod{n}$, which is impossible for $n \geqslant 3$. On the other hand, a symmetry σ is determined by $\sigma(0)$, one of the n vertices, and $\sigma(1)$, one of the two neighbors of $\sigma(0)$: once two adjacent vertices are placed, adjacency forces the images of the others, going around the polygon. So there are at most $2n$ symmetries, and the list is complete. The r^m are the n rotations. The n-gon also has n reflection symmetries, across the lines through the center and a vertex or the midpoint of an edge; so the reflections are the remaining elements sr^m. □

For $n = 3$, the 6 symmetries realize all $3! = 6$ permutations of the vertices, as in Example 2.2.13. For $n \geqslant 4$ we have $2n < n!$, so most permutations of the vertices do not come from symmetries. Since $r^{-1} \neq r$ for $n \geqslant 3$, the relation $sr = r^{-1}s$ shows that $sr \neq rs$: every dihedral group is non-abelian.

Example 2.2.16 (Symmetries of the square)

The group D_4 of symmetries of the square has 8 elements: the rotations e, r, r^2, r^3 by $0, \pi/2, \pi$, and $3\pi/2$, and four reflections.



With the vertices numbered as shown and s the reflection across the diagonal through vertices 0 and 2, the formula $sr^m(j) = -j - m$ identifies the reflections: sr swaps $0 \leftrightarrow 3$ and $1 \leftrightarrow 2$, so it is the reflection across the horizontal axis; sr^2 is the reflection across the diagonal through vertices 1 and 3; and sr^3 swaps $0 \leftrightarrow 1$ and $2 \leftrightarrow 3$, so it is the reflection across the vertical axis. In particular $rs = sr^{-1} = sr^3$ and sr are the reflections across the two axes, which differ.

2.2.5 Matrix Groups

Definition 2.2.17 (General linear group)

The ***general linear group*** $GL_n(\mathbb{R})$ is the group of invertible $n \times n$ real matrices under matrix multiplication. It is a group: matrix multiplication is associative, the identity matrix I is an identity, and a product of invertible

matrices is invertible, with $(AB)^{-1} = B^{-1}A^{-1}$. In the same way we have $GL_n(\mathbb{C})$ and $GL_n(\mathbb{Q})$; once fields are defined (Definition 4.1.6), $GL_n(k)$ makes sense for any field k.

Inside $GL_n(k)$ sits the ***special linear group*** $SL_n(k)$ of matrices with determinant 1. Determinants are the subject of Section 5.4, where we prove that a matrix is invertible if and only if its determinant is nonzero (Theorem 5.4.8) and that $\det(AB) = \det(A)\det(B)$ (Theorem 5.4.7); so $SL_n(k)$ is a subgroup of $GL_n(k)$. Until then SL_n appears only in examples.

NOTE Matrix groups over $\mathbb{R}$ and $\mathbb{C}$ are our first examples of *Lie groups*—groups that are also smooth manifolds. (Over a finite field such as $\mathbb{Z}/p\mathbb{Z}$, by contrast, GL_n is a finite group.) They are, in my opinion, the most exciting part of representation theory: a representation of G is a homomorphism $\rho : G \to GL_n(k)$ (Definition 13.1.1; homomorphisms are defined in Section 2.5), making G act on k^n by linear transformations.

Example 2.2.18 (A non-abelian matrix group)

In $GL_2(\mathbb{R})$, let $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $B = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$. Then

$$AB = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \neq \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} = BA.$$

So $GL_2(\mathbb{R})$ is non-abelian.

2.3 Constructions: Products and Sums

How do we build new groups from existing ones?

2.3.1 Direct Products

Definition 2.3.1 (Direct product)

Given groups G and H, the ***direct product*** $G \times H$ is the Cartesian product with the componentwise operation

$$(a, b) \cdot (a', b') = (aa', bb').$$

Its identity is (e_G, e_H), and the inverse of (a, b) is (a^{-1}, b^{-1}).

Proposition 2.3.2 (Direct product is a group)

The direct product $G \times H$ satisfies the group axioms. Moreover, if G and H are finite, then $|G \times H| = |G| \cdot |H|$.

Proof. **Associativity:** We have

$$((a,b)(a',b'))(a'',b'') = (aa',bb')(a'',b'') = ((aa')a'', (bb')b''),$$
$$(a,b)((a',b')(a'',b'')) = (a,b)(a'a'',b'b'') = (a(a'a''), b(b'b'')),$$

and these are equal by associativity in G and H.

Identity: $(e_G, e_H)(a, b) = (e_G a, e_H b) = (a, b)$, and similarly on the right.

Inverses: $(a, b)(a^{-1}, b^{-1}) = (aa^{-1}, bb^{-1}) = (e_G, e_H)$, and similarly on the other side.

Finally, $G \times H$ consists of the pairs (a, b), with $|G|$ choices for a and $|H|$ for b. □

The same construction works for any finite number of factors: $G_1 \times \cdots \times G_n$ is the set of n-tuples with componentwise operation. For example, $\mathbb{Z}^n = \mathbb{Z} \times \cdots \times \mathbb{Z}$ is the group of n-tuples of integers under componentwise addition, $(a_1, \ldots, a_n) + (b_1, \ldots, b_n) = (a_1 + b_1, \ldots, a_n + b_n)$; similarly for $\mathbb{Q}^n$, $\mathbb{R}^n$, and $\mathbb{C}^n$.

2.3.2 Infinite Products and Direct Sums

Products generalize to arbitrary (even infinite) families of groups.

Definition 2.3.3 (Arbitrary direct products)

For a family of groups $\{G_i\}_{i \in I}$, the ***direct product*** is the product of the underlying sets (Definition 1.4.6),

$$\prod_{i \in I} G_i = \{(a_i)_{i \in I} : a_i \in G_i \text{ for all } i\},$$

with the componentwise operation $(a_i)_{i \in I}(b_i)_{i \in I} = (a_i b_i)_{i \in I}$. As for two factors, it is a group, with identity $(e_{G_i})_{i \in I}$ and inverse $(a_i^{-1})_{i \in I}$.

Definition 2.3.4 (Direct sum)

For a family $\{G_i\}_{i\in I}$ of abelian groups, the ***direct sum*** is the subset

$$\bigoplus_{i\in I} G_i = \Big\{(a_i)_{i\in I} \in \prod_{i\in I} G_i : a_i = e_{G_i} \text{ for all but finitely many } i\Big\}.$$

It contains the identity and is closed under the operation and under inverses, since these act componentwise and preserve the condition of having only finitely many nonidentity terms. So it is a group in its own right, a subgroup of the product in the language of Section 2.4.

NOTE For finite index sets, $\prod = \bigoplus$. For infinite index sets, they differ crucially. The $\bigoplus$ notation will become important when we move on to linear algebra.

Example 2.3.5 (Power series and polynomials)

Take $G_i = (\mathbb{R}, +)$ for all $i \in \mathbb{N}$, and write the family $(a_0, a_1, a_2, \ldots)$ as $\sum_{i\geqslant 0} a_i x^i$. Then

$$\prod_{i\in\mathbb{N}} \mathbb{R} = \mathbb{R}[\![x]\!] \quad \text{(formal power series under addition)},$$
$$\bigoplus_{i\in\mathbb{N}} \mathbb{R} = \mathbb{R}[x] \quad \text{(polynomials under addition)}.$$

A power series $\sum_{i=0}^{\infty} a_i x^i$ can have infinitely many nonzero terms; a polynomial cannot. (Polynomials and power series can also be multiplied; that structure comes later, with rings.)

2.4 Subgroups

Definition 2.4.1 (Subgroup)

A ***subgroup*** of a group G is a *nonempty* subset $H \subseteq G$ such that:

(a) **Closure:** $a, b \in H \implies ab \in H$;

(b) **Inverses:** $a \in H \implies a^{-1} \in H$.

We write $H \leqslant G$ (or $H < G$ if $H \neq G$, a ***proper subgroup***).

If $H \leqslant G$ and $a \in H$, then $e = aa^{-1} \in H$. So H, with the operation of G restricted to it, is a group in its own right; associativity holds in H because it holds in G. Every group G has the ***trivial subgroup*** $\{e\}$ and the subgroup G itself.

Example 2.4.2 (First examples of subgroups)

(a) $(\mathbb{Z}, +) \leqslant (\mathbb{Q}, +) \leqslant (\mathbb{R}, +) \leqslant (\mathbb{C}, +)$, and $(\mathbb{Q}^*, \times) \leqslant (\mathbb{R}^*, \times) \leqslant (\mathbb{C}^*, \times) \geqslant (S^1, \times) \geqslant (\mu_n, \times)$.

(b) For each $n \geqslant 0$, the multiples of n, $n\mathbb{Z} = \{\ldots, -2n, -n, 0, n, 2n, \ldots\}$, form a subgroup of $\mathbb{Z}$.

(c) The rotations $\{e, r, \ldots, r^{n-1}\}$ form a subgroup of D_n.

(d) If $H_i \leqslant G_i$ for $i = 1, \ldots, n$, then $H_1 \times \cdots \times H_n \leqslant G_1 \times \cdots \times G_n$. For a family of abelian groups, $\bigoplus_i G_i \leqslant \prod_i G_i$.

(e) $SL_n(k) \leqslant GL_n(k)$, granting the multiplicativity of the determinant (Theorem 5.4.7).

Proposition 2.4.3 (Subgroup criterion)

A nonempty subset $H \subseteq G$ is a subgroup iff $a, b \in H \implies ab^{-1} \in H$.

Proof. ($\Rightarrow$) If $H \leqslant G$ and $a, b \in H$, then $b^{-1} \in H$ (inverses) and $ab^{-1} \in H$ (closure).

($\Leftarrow$) Assume the condition. Since $H \neq \emptyset$, pick $a \in H$. Then $e = aa^{-1} \in H$. For any $b \in H$, $b^{-1} = eb^{-1} \in H$. For $a, b \in H$, we have $b^{-1} \in H$, so $ab = a(b^{-1})^{-1} \in H$. □

Proposition 2.4.4 (Intersections of subgroups)

Let $\{H_i\}_{i \in I}$ be a nonempty family of subgroups of G. Then $\bigcap_{i \in I} H_i$ is a subgroup of G. In particular, if H and H' are subgroups of G, so is $H \cap H'$.

Proof. Each H_i contains e, so the intersection is nonempty. If a and b lie in every H_i, then so do ab and a^{-1}, since each H_i is a subgroup; so ab and a^{-1} lie in the intersection. □

Unions, on the other hand, are almost never subgroups: $H \cup H'$ is a subgroup only when one of H, H' contains the other. Indeed, suppose neither contains the other, and pick $h \in H \setminus H'$ and $h' \in H' \setminus H$. If $hh' \in H$, then $h' = h^{-1}(hh') \in H$, a contradiction; and if $hh' \in H'$, then $h = (hh')h'^{-1} \in H'$, also a contradiction. So $hh' \notin H \cup H'$. For example, $2\mathbb{Z} \cup 3\mathbb{Z}$ contains 2 and 3 but not 5, so it is not a subgroup of $\mathbb{Z}$, while $2\mathbb{Z} \cap 3\mathbb{Z} = 6\mathbb{Z}$ is a subgroup.

Theorem 2.4.5 (Subgroups of $\mathbb{Z}$)

Let H be a subgroup of $(\mathbb{Z}, +)$. Then either $H = \{0\}$ or $H = n\mathbb{Z}$, where n is the smallest positive element of H.

Proof. If $H = \{0\}$, we're done. Otherwise, H contains some nonzero element a; then $-a \in H$ too, so H contains a positive element. Let n be the smallest positive element of H.

We show $H = n\mathbb{Z}$. First, $n\mathbb{Z} \subseteq H$, since H is closed under addition and negation: $n \in H \implies 2n = n + n \in H \implies 3n \in H \implies \cdots$, and $-n \in H \implies -2n \in H \implies \cdots$.

For $H \subseteq n\mathbb{Z}$: take any $m \in H$. By the division algorithm, $m = qn + r$ with $0 \leqslant r < n$. Then $r = m - qn \in H$ (since $m, qn \in H$). By minimality of n, we must have $r = 0$, so $m = qn \in n\mathbb{Z}$. □

So the subgroups of $\mathbb{Z}$ are exactly the $n\mathbb{Z}$ with $n \geqslant 0$ (where $0\mathbb{Z} = \{0\}$). Some of them form chains, $\mathbb{Z} = 1\mathbb{Z} \supset 2\mathbb{Z} \supset 4\mathbb{Z} \supset 8\mathbb{Z} \supset \cdots$, but not all subgroups are comparable: neither of $2\mathbb{Z}$ and $3\mathbb{Z}$ contains the other. The theorem gives a quick route to the basic facts about greatest common divisors.

Proposition 2.4.6 (Greatest common divisor)

Let $a, b \in \mathbb{Z}$, not both zero, and let $d = \gcd(a, b)$ be their greatest common divisor. Then:

(a) $\mathbb{Z}a + \mathbb{Z}b = \{ra + sb : r, s \in \mathbb{Z}\} = d\mathbb{Z}$;

(b) if n divides both a and b, then n divides d;

(c) there exist $r, s \in \mathbb{Z}$ such that $d = ra + sb$ (***Bézout's identity***).

Proof. The set $S = \mathbb{Z}a + \mathbb{Z}b$ is a subgroup of $\mathbb{Z}$: it contains 0, and $(ra + sb) - (r'a + s'b) = (r - r')a + (s - s')b$. It contains a and b, not both zero, so $S \neq \{0\}$, and by Theorem 2.4.5, $S = d'\mathbb{Z}$, where d' is the smallest positive element of S. Since $a = 1 \cdot a + 0 \cdot b$ and b lie in $S = d'\mathbb{Z}$, d' divides both a and b. Since $d' \in S$, we have $d' = ra + sb$ for some r, s; so if n divides a and b, then n divides $ra + sb = d'$, and in particular $|n| \leqslant d'$. Thus d' is a common divisor of a and b that is at least as large as every other one, that is, $d' = \gcd(a, b) = d$. This proves (a) and (c), and (b) was shown along the way. □

Corollary 2.4.7 (Coprimality criterion)

Integers a and b are coprime (i.e., $\gcd(a, b) = 1$) iff $\exists r, s \in \mathbb{Z}$ with $ra + sb = 1$.

Proof. If $\gcd(a,b)=1$, this is Bézout's identity. Conversely, if $ra+sb=1$, every common divisor of a and b divides 1, so $\gcd(a,b)=1$. □

Corollary 2.4.8 (Euclid's lemma)

Let n,k,m be integers with $n \mid km$ and $\gcd(n,k)=1$. Then $n \mid m$. In particular, if p is prime and $p \mid ab$, then $p \mid a$ or $p \mid b$.

Proof. By Bézout's identity, $rn+sk=1$ for some $r,s\in\mathbb{Z}$. Multiplying by m gives $m=rnm+skm$, and n divides both terms on the right, so $n \mid m$. For the second statement, suppose $p \nmid a$. Since the only positive divisors of p are 1 and p, we have $\gcd(p,a)=1$, and the first statement gives $p \mid b$. □

As an application we obtain a multiplicative group of integers mod n. For $n \geqslant 2$ and $a,b \in \mathbb{Z}/n\mathbb{Z}$, let $a \cdot_n b$ be the remainder of ab on division by n (***multiplication mod*** n). As for addition, $(a \cdot_n b) \cdot_n c$ and $a \cdot_n (b \cdot_n c)$ are both the remainder of abc, so multiplication mod n is associative, and 1 is an identity for it.

Proposition 2.4.9 (Units mod n)

Let $n \geqslant 2$. The set $(\mathbb{Z}/n\mathbb{Z})^\times = \{a \in \mathbb{Z}/n\mathbb{Z} : \gcd(a,n)=1\}$ forms a group under multiplication mod n. Its order is $\varphi(n)$, the number of integers $1 \leqslant a \leqslant n$ coprime to n (***Euler's totient function***).

Proof. **Closure:** Let $\gcd(a,n)=\gcd(b,n)=1$, and let $c = a \cdot_n b$, so that $ab = qn + c$ for some q. If a prime p divided both c and n, it would divide ab, hence a or b by Euclid's lemma, contradicting coprimality. Since every integer greater than 1 has a prime divisor, $\gcd(c,n)=1$.

Identity: 1 is coprime to n.

Inverses: If $\gcd(a,n)=1$, then by Bézout's identity there are r,s with $ra+sn=1$. Write $r = tn + r'$ with $0 \leqslant r' \leqslant n-1$. Then $r'a + (s+ta)n = 1$, so r' is coprime to n by Corollary 2.4.7, and $r'a$ is one more than a multiple of n, that is, $r' \cdot_n a = 1$. So $r' \in (\mathbb{Z}/n\mathbb{Z})^\times$ is an inverse of a.

The elements of $(\mathbb{Z}/n\mathbb{Z})^\times$ are the integers $1 \leqslant a \leqslant n-1$ coprime to n (neither 0 nor n is coprime to $n \geqslant 2$), so there are $\varphi(n)$ of them. □

2.4.1 Generated Subgroups and Cyclic Groups

Given a subset $S \subseteq G$, what is the smallest subgroup of G containing S? There is at least one subgroup containing S, namely G itself, and by Proposition 2.4.4 the intersection of all of them is again a subgroup.

Definition 2.4.10 (Subgroup generated by a set)

For $S \subseteq G$, the ***subgroup generated by*** S is

$$\langle S \rangle = \bigcap_{\substack{H \leqslant G \\ S \subseteq H}} H.$$

It is a subgroup of G that contains S and is contained in every subgroup containing S; so it is the smallest subgroup of G containing S. If $\langle S \rangle = G$, we say that S ***generates*** G. For $S = \{a_1, \ldots, a_k\}$ we write $\langle a_1, \ldots, a_k \rangle$.

Proposition 2.4.11 (Elements of a generated subgroup)

For $S \subseteq G$,

$$\langle S \rangle = \{s_1^{\epsilon_1} s_2^{\epsilon_2} \cdots s_k^{\epsilon_k} : k \geqslant 0,\ s_i \in S,\ \epsilon_i \in \{1, -1\}\},$$

where the empty product ($k = 0$) is e. In particular, for $a \in G$, $\langle a \rangle = \{a^n \mid n \in \mathbb{Z}\}$.

Proof. Let W be the set on the right. It contains e; the product of two such words is again a word; and the inverse of a word is a word, since $(s_1^{\epsilon_1} \cdots s_k^{\epsilon_k})^{-1} = s_k^{-\epsilon_k} \cdots s_1^{-\epsilon_1}$ by Proposition 2.1.5 and induction on k. So W is a subgroup containing S, and hence $\langle S \rangle \subseteq W$. Conversely, a subgroup containing S contains every $s^{\pm 1}$ with $s \in S$ and all products of these, so it contains W; hence $W \subseteq \langle S \rangle$. When $S = \{a\}$, a word in a and a^{-1} equals a^n, where n is the number of factors a minus the number of factors a^{-1}. □

Definition 2.4.12 (Cyclic groups)

A group G is ***cyclic*** if $G = \langle a \rangle$ for some $a \in G$. The element a is called a ***generator***. For any $a \in G$, the subgroup $\langle a \rangle = \{a^n \mid n \in \mathbb{Z}\}$ is the ***cyclic subgroup*** generated by a.

Example 2.4.13 (Cyclic groups and generators)

$\mathbb{Z} = \langle 1 \rangle = \langle -1 \rangle$ and $\mathbb{Z}/n\mathbb{Z} = \langle 1 \rangle$ are cyclic, as is $\mu_n = \langle e^{2\pi i/n} \rangle$. By Theorem 2.4.5, every subgroup of $\mathbb{Z}$ is cyclic: $n\mathbb{Z} = \langle n \rangle$. The rotation subgroup of D_n is $\langle r \rangle$, and $D_n = \langle r, s \rangle$ by Proposition 2.2.15. By contrast, $SL_2(\mathbb{Z})$, the group of integer matrices $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ with $ad - bc = 1$ under multiplication, is not cyclic (it is not even abelian: the matrices A and B of

Example 2.2.18 lie in it and do not commute), but it can be generated by two elements (a fairly hard exercise without a hint; we return to this group in Section 12.4).

Definition 2.4.14 (Order of an element)

The ***order*** of $a \in G$, denoted $\operatorname{ord}(a)$ or $|a|$, is the smallest positive integer n such that $a^n = e$, or ∞ if no such n exists; in the latter case we say that a has ***infinite order***.

Do not confuse the order of an element with the order $|G|$ of a group. The two are related: $\operatorname{ord}(a) = |\langle a \rangle|$, as we prove in Proposition 2.6.1.

2.5 Homomorphisms

Definition 2.5.1 (Homomorphism)

A ***homomorphism*** $\varphi : G \to H$ between groups is a map satisfying

$$\varphi(ab) = \varphi(a)\varphi(b) \quad \text{for all } a, b \in G.$$

Proposition 2.5.2 (Basic properties of homomorphisms)

Let $\varphi : G \to H$ be a homomorphism.

(a) $\varphi(e_G) = e_H$.
(b) $\varphi(a^{-1}) = \varphi(a)^{-1}$ for all $a \in G$.
(c) $\varphi(a^n) = \varphi(a)^n$ for all $a \in G$ and $n \in \mathbb{Z}$.
(d) If $K \leqslant G$, then $\varphi(K) \leqslant H$.
(e) If $L \leqslant H$, then $\varphi^{-1}(L) \leqslant G$.

Proof. **(a)** $\varphi(e_G) = \varphi(e_G e_G) = \varphi(e_G)\varphi(e_G)$. Multiply both sides by $\varphi(e_G)^{-1}$: $e_H = \varphi(e_G)$.

(b) $\varphi(a)\varphi(a^{-1}) = \varphi(aa^{-1}) = \varphi(e_G) = e_H$, and similarly $\varphi(a^{-1})\varphi(a) = e_H$. So $\varphi(a^{-1})$ is the inverse of $\varphi(a)$.

(c) For $n > 0$: induction, using $\varphi(a^n) = \varphi(a^{n-1})\varphi(a)$. For $n = 0$: part (a). For $n < 0$: $\varphi(a^n) = \varphi((a^{-1})^{-n}) = \varphi(a^{-1})^{-n} = (\varphi(a)^{-1})^{-n} = \varphi(a)^n$.

(d) The set $\varphi(K)$ is nonempty, since K is. For $\varphi(a), \varphi(b) \in \varphi(K)$ (with $a, b \in K$): $\varphi(a)\varphi(b)^{-1} = \varphi(a)\varphi(b^{-1}) = \varphi(ab^{-1}) \in \varphi(K)$, since $ab^{-1} \in K$. Apply the subgroup criterion.

(e) The set $\varphi^{-1}(L)$ contains e_G, since $\varphi(e_G) = e_H \in L$. For $a, b \in \varphi^{-1}(L)$: $\varphi(ab^{-1}) = \varphi(a)\varphi(b)^{-1} \in L$ (since $L \leqslant H$), so $ab^{-1} \in \varphi^{-1}(L)$. Apply the subgroup criterion. □

NOTE The homomorphism condition can be expressed as a *commutative diagram*: the two paths $G \times G \xrightarrow{m_G} G \xrightarrow{\varphi} H$ and $G \times G \xrightarrow{\varphi\times\varphi} H \times H \xrightarrow{m_H} H$ give the same result. It does not matter whether we multiply first or apply φ first.

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

Definition 2.5.3 (Types of homomorphisms)

Let $\varphi : G \to H$ be a homomorphism.

(a) φ is a ***monomorphism*** (or ***embedding***) if it is injective.

(b) φ is an ***epimorphism*** if it is surjective.

(c) φ is an ***isomorphism*** if it is bijective.

(d) An isomorphism $G \to G$ is called an ***automorphism*** of G. The set of all automorphisms of G is denoted $\mathrm{Aut}(G)$.

Proposition 2.5.4 (Composites and inverses)

Let $\varphi : G \to H$ and $\psi : H \to K$ be homomorphisms.

(a) $\psi \circ \varphi : G \to K$ is a homomorphism, and if φ and ψ are isomorphisms, so is $\psi \circ \varphi$.

(b) If φ is an isomorphism, then so is its inverse $\varphi^{-1} : H \to G$.

(c) $\mathrm{Aut}(G)$ is a group under composition; it is a subgroup of $\mathrm{Sym}(G)$.

Proof. (a) $\psi(\varphi(ab)) = \psi(\varphi(a)\varphi(b)) = \psi(\varphi(a))\psi(\varphi(b))$, and a composite of bijections is a bijection.

(b) φ^{-1} is a bijection (Proposition 1.5.5). For $h_1, h_2 \in H$, let $g_i = \varphi^{-1}(h_i)$. Then $\varphi^{-1}(h_1h_2) = \varphi^{-1}(\varphi(g_1)\varphi(g_2)) = \varphi^{-1}(\varphi(g_1g_2)) = g_1g_2 = \varphi^{-1}(h_1)\varphi^{-1}(h_2)$.

(c) $\mathrm{Aut}(G) \subseteq \mathrm{Sym}(G)$ contains id_G, and it is closed under composition and inverses by (a) and (b). □

2.5.1 Kernel and Image

Definition 2.5.5 (Kernel and image)

Let $\varphi : G \to H$ be a homomorphism. The ***kernel*** and the ***image*** of φ are

$$\ker \varphi = \{a \in G \mid \varphi(a) = e_H\}, \qquad \mathrm{Im}\, \varphi = \varphi(G) = \{\varphi(a) \mid a \in G\}.$$

Proposition 2.5.6 (Properties of the kernel and image)

Let $\varphi : G \to H$ be a homomorphism.

(a) $\ker \varphi$ is a subgroup of G, and $\mathrm{Im}\, \varphi$ is a subgroup of H.
(b) For $a, b \in G$, $\varphi(a) = \varphi(b)$ if and only if $a^{-1}b \in \ker \varphi$.
(c) φ is injective if and only if $\ker \varphi = \{e_G\}$.
(d) φ is surjective if and only if $\mathrm{Im}\, \varphi = H$.
(e) If φ is injective, then $a \mapsto \varphi(a)$ is an isomorphism $G \to \mathrm{Im}\, \varphi$; so G is isomorphic to the subgroup $\mathrm{Im}\, \varphi$ of H.

Proof. (a) $\ker \varphi = \varphi^{-1}(\{e_H\})$ and $\mathrm{Im}\, \varphi = \varphi(G)$, so both are subgroups by Proposition 2.5.2(d) and (e).

(b) $\varphi(a) = \varphi(b) \iff \varphi(a)^{-1}\varphi(b) = e_H \iff \varphi(a^{-1}b) = e_H \iff a^{-1}b \in \ker \varphi$.

(c) If φ is injective and $a \in \ker \varphi$, then $\varphi(a) = e_H = \varphi(e_G)$, so $a = e_G$. Conversely, if $\ker \varphi = \{e_G\}$ and $\varphi(a) = \varphi(b)$, then $a^{-1}b = e_G$ by (b), so $a = b$.

(d) This is the definition of surjectivity.

(e) The map $G \to \mathrm{Im}\, \varphi$, $a \mapsto \varphi(a)$, is a homomorphism, injective because φ is, and surjective by definition of $\mathrm{Im}\, \varphi$. □

Example 2.5.7 (Important(ish) homomorphisms)

(a) **Inclusion:** if $H \leqslant G$, the inclusion map $H \hookrightarrow G$ is an injective homomorphism.
(b) **Reduction mod n:** $\pi : \mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$, where $\pi(a)$ is the remainder of a on division by n. Both $\pi(a + b)$ and $\pi(a) +_n \pi(b)$ are the remainder of $a + b$, so π is a homomorphism. It is surjective, with kernel $n\mathbb{Z}$.
(c) **Reduction from $\mathbb{Z}/m\mathbb{Z}$ to $\mathbb{Z}/n\mathbb{Z}$:** if $n \mid m$, then $\mathbb{Z}/m\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$, sending a to its remainder mod n, is a surjective homomorphism, because $a +_m b$

differs from $a + b$ by 0 or m, a multiple of n. Example: last two digits $\mapsto$ last digit, $\mathbb{Z}/100\mathbb{Z} \to \mathbb{Z}/10\mathbb{Z}$.

(d) **Powers of an element:** for $a \in G$, the map $\mathbb{Z} \to G$, $n \mapsto a^n$, is a homomorphism, since $a^{m+n} = a^m a^n$. By Proposition 2.4.11, its image is the cyclic subgroup $\langle a \rangle$.

(e) **Determinant:** the determinant of Section 5.4, $\det : \mathrm{GL}_n(\mathbb{R}) \to \mathbb{R}^*$, is a homomorphism, since $\det(AB) = \det(A)\det(B)$ (Theorem 5.4.7). Its kernel is $\mathrm{SL}_n(\mathbb{R})$.

(f) **Exponential:** $\exp : (\mathbb{R}, +) \to (\mathbb{R}_{>0}, \times)$, $x \mapsto e^x$, is an isomorphism, with inverse $\ln$.

(g) **Complex exponential:** $\mathbb{R} \to S^1$, $t \mapsto e^{2\pi i t}$, is a surjective homomorphism from $(\mathbb{R}, +)$ to $(S^1, \times)$ with kernel $\mathbb{Z}$. Restricted to $[0, 1)$, it gives an isomorphism $\mathbb{R}/\mathbb{Z} \to S^1$: every $z \in S^1$ equals $e^{2\pi i t}$ for exactly one $t \in [0, 1)$, and since $a +_1 b$ differs from $a + b$ by an integer m and $e^{2\pi i m} = 1$, we have $e^{2\pi i (a +_1 b)} = e^{2\pi i a} e^{2\pi i b}$.

2.5.2 Isomorphisms and Invariants

Definition 2.5.8 (Isomorphic groups)

Groups G and H are ***isomorphic***, written $G \cong H$, if there exists an isomorphism $\varphi : G \to H$. Isomorphic groups are "the same" for all group-theoretic purposes: the elements and the law may have different names, but the structure is identical.

What properties are preserved by isomorphisms? These are called *invariants*.

Proposition 2.5.9 (Isomorphism invariants)

If $G \cong H$, then:

(a) $|G| = |H|$ (same order);

(b) G is abelian iff H is abelian;

(c) G is cyclic iff H is cyclic;

(d) for each n, G and H have the same number of elements of order n;

(e) for each n, G and H have the same number of subgroups of order n.

Proof. Let $\varphi : G \to H$ be an isomorphism. Since $\varphi^{-1} : H \to G$ is also an isomorphism (Proposition 2.5.4), for (b) and (c) it suffices to prove one direction.

(a) φ is a bijection.

(b) If G is abelian and $h_1, h_2 \in H$, write $h_i = \varphi(g_i)$. Then $h_1h_2 = \varphi(g_1)\varphi(g_2) = \varphi(g_1g_2) = \varphi(g_2g_1) = \varphi(g_2)\varphi(g_1) = h_2h_1$.

(c) If $G = \langle a \rangle = \{a^n \mid n \in \mathbb{Z}\}$, then $H = \varphi(G) = \{\varphi(a)^n \mid n \in \mathbb{Z}\} = \langle \varphi(a) \rangle$, using Proposition 2.5.2(c).

(d) For $a \in G$ and $n \in \mathbb{Z}$, $\varphi(a)^n = \varphi(a^n)$, which equals $e_H = \varphi(e_G)$ if and only if $a^n = e_G$, since φ is injective. So $\operatorname{ord}(\varphi(a)) = \operatorname{ord}(a)$, and φ restricts to a bijection from the elements of order n in G to those in H.

(e) By Proposition 2.5.2, $K \mapsto \varphi(K)$ sends subgroups of G to subgroups of H, and $L \mapsto \varphi^{-1}(L)$ sends subgroups of H to subgroups of G. Since φ is bijective, $\varphi^{-1}(\varphi(K)) = K$ and $\varphi(\varphi^{-1}(L)) = L$ (Proposition 1.5.2), so these are inverse bijections between the subgroups of G and those of H; and $|\varphi(K)| = |K|$, since φ is injective. □

Example 2.5.10 (Non-isomorphic groups of the same order)

$\mathbb{Z}/4\mathbb{Z}$ and $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ both have order 4, but:

- $\mathbb{Z}/4\mathbb{Z}$ has an element of order 4 (namely, 1);
- $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ has no element of order 4, since $x + x = 0$ for every x (all non-identity elements have order 2).

So they are not isomorphic. In Question 2.6.10 we show that these are the only groups of order 4 up to isomorphism.

Example 2.5.11 (All groups of order 2 are isomorphic)

If $|G| = 2$, write $G = \{e, x\}$ with $x \neq e$. If $x^2 = x$, then cancellation would give $x = e$; so $x^2 = e$, and the multiplication table of G is

	e	x
e	e	x
x	x	e

Any two groups of order 2 have this table up to renaming, so the bijection matching their identities and their other elements is an isomorphism. For example, $S_2 = \{e, (1\,2)\}$, $(\{\pm 1\}, \times)$, and $(\mathbb{Z}/2\mathbb{Z}, +)$ are all isomorphic.

Similarly, sending a symmetry of the triangle to the permutation of the vertices it induces is an isomorphism $D_3 \to S_3$ (Example 2.2.13).

Question 2.5.12

(a) Find $\operatorname{Aut}(\mathbb{Z})$.

(b) Prove that "is isomorphic to" is an equivalence relation on groups.

Solution. **(a)** An automorphism φ of $\mathbb{Z}$ is determined by $\varphi(1)$, since $\varphi(k) = k\varphi(1)$ by Proposition 2.5.2(c). Its image is $\{k\varphi(1) \mid k \in \mathbb{Z}\} = \langle\varphi(1)\rangle$, which must be all of $\mathbb{Z}$, so $\varphi(1) = \pm 1$. Both choices give automorphisms, id and $-\mathrm{id}$. Thus $\mathrm{Aut}(\mathbb{Z}) = \{\mathrm{id}, -\mathrm{id}\} \cong \mathbb{Z}/2\mathbb{Z}$ (Example 2.5.11).
(b) Reflexive: $\mathrm{id}_G : G \to G$ is an isomorphism. Symmetric: if $\varphi : G \to H$ is an isomorphism, so is $\varphi^{-1} : H \to G$. Transitive: if $\varphi : G \to H$ and $\psi : H \to K$ are isomorphisms, so is $\psi \circ \varphi : G \to K$. The last two statements are Proposition 2.5.4.

2.6 Orders of Elements and Cyclic Groups

We have introduced the order of an element; now we develop its theory more fully and use it to classify cyclic groups.

Proposition 2.6.1 (Powers and order)

Let $a \in G$.

(a) If a has finite order n, then $a^k = e \iff n \mid k$, and $a^i = a^j \iff n \mid (i-j)$. The elements $e, a, a^2, \ldots, a^{n-1}$ are all distinct, and $\langle a\rangle = \{e, a, \ldots, a^{n-1}\}$; so $|\langle a\rangle| = n$.

(b) If a has infinite order, then the powers a^k, $k \in \mathbb{Z}$, are pairwise distinct; so $\langle a\rangle$ is infinite.

In either case, $\mathrm{ord}(a) = |\langle a\rangle|$.

Proof. **(a)** ($\Leftarrow$) If $k = mn$, then $a^k = (a^n)^m = e^m = e$. ($\Rightarrow$) Write $k = qn + r$ with $0 \leqslant r < n$. Then $e = a^k = a^{qn+r} = (a^n)^q a^r = a^r$. By minimality of n, $r = 0$, so $n \mid k$. Next, $a^i = a^j \iff a^{i-j} = e \iff n \mid (i-j)$. If $a^i = a^j$ for $0 \leqslant i < j < n$, then n would divide $j - i$, with $0 < j - i < n$, which is impossible; so $e, a, \ldots, a^{n-1}$ are distinct. Finally, every power a^k equals a^r, where r is the remainder of k on division by n; so $\langle a\rangle = \{a^k \mid k \in \mathbb{Z}\} = \{e, a, \ldots, a^{n-1}\}$.

(b) If $a^i = a^j$ with $i > j$, then $a^{i-j} = e$ with $i - j > 0$, contradicting infinite order. □

Theorem 2.6.2 (Structure of cyclic groups)

Let $G = \langle a\rangle$.

(a) If $\mathrm{ord}(a) = \infty$, then $\mathbb{Z} \to G$, $k \mapsto a^k$, is an isomorphism, so $G \cong \mathbb{Z}$.

(b) If $\operatorname{ord}(a) = n < \infty$, then $G = \{e, a, a^2, \dots, a^{n-1}\}$, these n elements are distinct, and $\mathbb{Z}/n\mathbb{Z} \to G$, $k \mapsto a^k$ $(0 \leqslant k \leqslant n-1)$, is an isomorphism, so $G \cong \mathbb{Z}/n\mathbb{Z}$.

Thus every cyclic group is isomorphic to $\mathbb{Z}$ or to $\mathbb{Z}/n\mathbb{Z}$ for some $n \geqslant 1$.

Proof. The map $\varphi : \mathbb{Z} \to G$, $\varphi(k) = a^k$, is a homomorphism with image $\langle a \rangle = G$ (Example 2.5.7), so it is surjective.

(a) If $\operatorname{ord}(a) = \infty$, the powers of a are pairwise distinct by Proposition 2.6.1, so φ is injective, hence an isomorphism $\mathbb{Z} \xrightarrow{\sim} G$.

(b) If $\operatorname{ord}(a) = n$, Proposition 2.6.1 shows that $G = \{e, a, \dots, a^{n-1}\}$ with these elements distinct, so $\psi : \mathbb{Z}/n\mathbb{Z} \to G$, $\psi(k) = a^k$, is a bijection. It is a homomorphism: $k +_n l$ differs from $k+l$ by a multiple of n, so by Proposition 2.6.1, $\psi(k +_n l) = a^{k+_n l} = a^{k+l} = a^k a^l = \psi(k)\psi(l)$. □

NOTE In case (b), $\ker \varphi = n\mathbb{Z}$ by Proposition 2.6.1, and ψ is the map that φ induces on $\mathbb{Z}/n\mathbb{Z}$. In the next chapter this becomes an instance of the First Isomorphism Theorem (Theorem 3.3.3): $\mathbb{Z}/\ker\varphi \cong \operatorname{Im}\varphi$.

Example 2.6.3 (Cyclic groups in disguise)

(a) $\mu_n = \langle e^{2\pi i/n} \rangle$ has order n, so $\mu_n \cong \mathbb{Z}/n\mathbb{Z}$ via $k \mapsto e^{2\pi i k/n}$.

(b) The rotation subgroup $\langle r \rangle$ of D_n is isomorphic to $\mathbb{Z}/n\mathbb{Z}$, since r has order n.

(c) In $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$, the element $(1,1)$ satisfies $k(1,1) = (0,0)$ if and only if $2 \mid k$ and $3 \mid k$, that is, $6 \mid k$. So $(1,1)$ has order 6, the subgroup it generates has 6 elements, and $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z} = \langle (1,1) \rangle$. By Theorem 2.6.2, $k \mapsto k(1,1) = (k \bmod 2,\ k \bmod 3)$ is an isomorphism $\mathbb{Z}/6\mathbb{Z} \to \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$. On the other hand, $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z} \ncong \mathbb{Z}/4\mathbb{Z}$, since the former has no element of order 4.

Proposition 2.6.4 (Order of powers)

Let $\operatorname{ord}(a) = n < \infty$, and let $k \in \mathbb{Z}$. Then $\operatorname{ord}(a^k) = \dfrac{n}{\gcd(k, n)}$.

Proof. Let $d = \gcd(k, n)$, and write $k = dk'$ and $n = dn'$. Then $\gcd(k', n') = 1$: if $c > 0$ divides both k' and n', then cd divides both k and n, so $cd \leqslant d$ and $c = 1$.

First, $(a^k)^{n'} = a^{dk'n'} = (a^n)^{k'} = e$, so $\operatorname{ord}(a^k)$ is finite and divides n' by Proposition 2.6.1.

Conversely, let $m = \operatorname{ord}(a^k)$. Then $a^{km} = e$, so $n \mid km$, that is, $dn' \mid dk'm$, and so $n' \mid k'm$. Since $\gcd(k', n') = 1$, Euclid's lemma (Corollary 2.4.8) gives $n' \mid m$. As m and n' are positive and divide each other, $m = n' = n/d$. □

Corollary 2.6.5 (Generators of cyclic groups)

In a cyclic group $\langle a \rangle$ of order n, the element a^k is a generator iff $\gcd(k, n) = 1$.

Proof. Since $\langle a^k \rangle \subseteq \langle a \rangle$ and both are finite, $\langle a^k \rangle = \langle a \rangle$ if and only if $|\langle a^k \rangle| = n$, that is, $\operatorname{ord}(a^k) = n$ (Proposition 2.6.1). By Proposition 2.6.4, this happens if and only if $\gcd(k, n) = 1$. □

Example 2.6.6

In $\mathbb{Z}/12\mathbb{Z}$, the generators are $1, 5, 7, 11$ (the four elements coprime to 12). The element 4 has order $12/\gcd(4, 12) = 12/4 = 3$.

Proposition 2.6.7 (Subgroups of cyclic groups)

Every subgroup of a cyclic group is cyclic. More precisely, if $G = \langle a \rangle$ has order n, then the subgroups of G are exactly the subgroups $\langle a^d \rangle$ for the positive divisors d of n; these are distinct, and $|\langle a^d \rangle| = n/d$. So G has exactly one subgroup of each order dividing n.

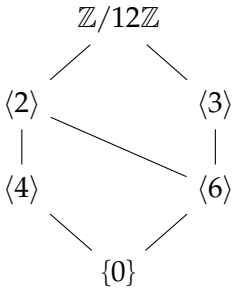
Proof. Let $H \leqslant \langle a \rangle$. If $H = \{e\}$, then $H = \langle e \rangle$ is cyclic. Otherwise H contains some $a^k \neq e$, and with it a^{-k}, so H contains a positive power of a. Let d be the smallest positive integer with $a^d \in H$. We claim $H = \langle a^d \rangle$. Clearly $\langle a^d \rangle \subseteq H$. Conversely, take $a^k \in H$, and write $k = qd + r$ with $0 \leqslant r < d$. Then $a^r = a^k (a^d)^{-q} \in H$. By minimality of d, $r = 0$, so $a^k = (a^d)^q \in \langle a^d \rangle$. So every subgroup of G is cyclic.

Now suppose $|G| = n$, so that $\operatorname{ord}(a) = n$ by Proposition 2.6.1. If $H \neq \{e\}$ and d is as above, then $a^n = e \in H$, and the same division argument, applied to $k = n$, shows that $d \mid n$. If $H = \{e\}$, then $H = \langle a^n \rangle$, and n divides n. So every subgroup of G is $\langle a^d \rangle$ for some positive divisor d of n. For such d, Propositions 2.6.1 and 2.6.4 give $|\langle a^d \rangle| = \operatorname{ord}(a^d) = n/\gcd(d, n) = n/d$. Different divisors d give different orders n/d, so the subgroups $\langle a^d \rangle$ are distinct, and as d runs over the divisors of n, so does n/d. □

Question 2.6.8

(a) Find all subgroups of $\mathbb{Z}/12\mathbb{Z}$ and draw the subgroup lattice.
(b) Prove that $\mathbb{Z}/p\mathbb{Z}$ has no proper nontrivial subgroups for p prime.
(c) How many generators does $\mathbb{Z}/n\mathbb{Z}$ have?
(d) Prove that $\mathrm{Aut}(\mathbb{Z}/n\mathbb{Z}) \cong (\mathbb{Z}/n\mathbb{Z})^\times$ for $n \geqslant 2$.
(e) Prove: if G is a group in which every element satisfies $x^2 = e$, then G is abelian.

Solution. **(a)** By Proposition 2.6.7, the subgroups are $\langle d \rangle$ for the divisors d of 12, namely $1, 2, 3, 4, 6, 12$: $\langle 1 \rangle = \mathbb{Z}/12\mathbb{Z}$, $\langle 2 \rangle = \{0, 2, 4, 6, 8, 10\}$, $\langle 3 \rangle = \{0, 3, 6, 9\}$, $\langle 4 \rangle = \{0, 4, 8\}$, $\langle 6 \rangle = \{0, 6\}$, and $\langle 12 \rangle = \{0\}$. Moreover, $\langle d \rangle \subseteq \langle d' \rangle$ exactly when d' divides d: if $d = d'c$, then $d = c \cdot d' \in \langle d' \rangle$; conversely, the elements of $\langle d' \rangle$ are the multiples of d' in $\{0, \ldots, 11\}$ (because $d' \mid 12$), so $d \in \langle d' \rangle$ forces $d' \mid d$.
Lattice (larger subgroups higher, an edge for each inclusion with no subgroup in between):



Note that $\langle 4 \rangle \not\subseteq \langle 3 \rangle$, since 4 is not a multiple of 3.
(b) By Proposition 2.6.7, the subgroups of $\mathbb{Z}/p\mathbb{Z} = \langle 1 \rangle$ correspond to the positive divisors of p, which are 1 and p. They are $\langle 1 \rangle = \mathbb{Z}/p\mathbb{Z}$ and $\langle p \rangle = \{0\}$.
(c) The element $k \in \mathbb{Z}/n\mathbb{Z}$ is $k \cdot 1$, the kth power of the generator 1 in additive notation. By Corollary 2.6.5, it generates $\mathbb{Z}/n\mathbb{Z}$ iff $\gcd(k, n) = 1$. The count is $\varphi(n)$, Euler's totient.
(d) An automorphism φ of $\mathbb{Z}/n\mathbb{Z}$ is determined by $\varphi(1)$, since $\varphi(k) = k\varphi(1)$. Its image $\langle \varphi(1) \rangle$ must be all of $\mathbb{Z}/n\mathbb{Z}$, so $\varphi(1)$ is a generator, i.e., $\gcd(\varphi(1), n) = 1$ by (c), i.e., $\varphi(1) \in (\mathbb{Z}/n\mathbb{Z})^\times$. Conversely, for $a \in (\mathbb{Z}/n\mathbb{Z})^\times$, let $\varphi_a : k \mapsto k \cdot_n a$. It is a homomorphism, since $(k +_n l)a$ and $ka + la$ differ by a multiple of n; and if b is the inverse of a in $(\mathbb{Z}/n\mathbb{Z})^\times$, then $\varphi_b \circ \varphi_a$ and $\varphi_a \circ \varphi_b$ are multiplication by $ab \equiv 1$, the identity, so φ_a is an automorphism. The map $a \mapsto \varphi_a$ satisfies $\varphi_{a \cdot_n b} = \varphi_a \circ \varphi_b$, so it is a

homomorphism $(\mathbb{Z}/n\mathbb{Z})^\times \to \mathrm{Aut}(\mathbb{Z}/n\mathbb{Z})$. It is injective, since $\varphi_a(1) = a$, and surjective by the first part ($\varphi = \varphi_{\varphi(1)}$). So it is an isomorphism.
(e) For any $a, b \in G$: $ab = (ab)^{-1} = b^{-1}a^{-1} = ba$ (using $x^{-1} = x$, since $x^2 = e$). So G is abelian.

Question 2.6.9

(a) In a group G, if $a^{12} = e$ and $a^8 = e$, what are the possible orders of a?
(b) Prove: if $\mathrm{ord}(a) = n$ and $\mathrm{ord}(b) = m$ with $\gcd(n, m) = 1$ and $ab = ba$, then $\mathrm{ord}(ab) = nm$.

Solution. **(a)** By Proposition 2.6.1, $\mathrm{ord}(a)$ divides both 12 and 8, so $\mathrm{ord}(a) \mid \gcd(12, 8) = 4$ by Proposition 2.4.6. Possible orders: 1, 2, 4. All three occur, for instance for 0, 2, and 1 in $\mathbb{Z}/4\mathbb{Z}$.
(b) Let $k = \mathrm{ord}(ab)$. Since a and b commute, $(ab)^{nm} = a^{nm}b^{nm} = (a^n)^m(b^m)^n = e$, so k is finite and $k \mid nm$.
Now $e = (ab)^{kn} = a^{kn}b^{kn} = b^{kn}$ (since $a^n = e$), so $m \mid kn$. Since $\gcd(m, n) = 1$, Euclid's lemma gives $m \mid k$. Similarly, $n \mid k$. Writing $k = mk'$, we have $n \mid mk'$, so $n \mid k'$ by Euclid's lemma again, and $nm \mid k$. Thus $k = nm$.

Question 2.6.10

(a) Prove that the order of $(a, b) \in G \times H$ is the least common multiple of $\mathrm{ord}(a)$ and $\mathrm{ord}(b)$ (when both are finite), and find the order of $(2, 3)$ in $\mathbb{Z}/6\mathbb{Z} \times \mathbb{Z}/8\mathbb{Z}$.
(b) Prove that $\mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}/mn\mathbb{Z}$ if and only if $\gcd(m, n) = 1$.
(c) Find all groups of order 4 up to isomorphism.
(d) Prove that $G \times H \cong H \times G$ for any groups G, H.

Solution. **(a)** Let $\mathrm{ord}(a) = r$ and $\mathrm{ord}(b) = s$. Since $(a, b)^k = (a^k, b^k)$, Proposition 2.6.1 gives $(a, b)^k = e$ if and only if $r \mid k$ and $s \mid k$, that is, $k \in r\mathbb{Z} \cap s\mathbb{Z}$. This intersection is a subgroup of $\mathbb{Z}$ (Proposition 2.4.4) containing $rs \neq 0$, so by Theorem 2.4.5 it is $L\mathbb{Z}$, where L is its smallest positive element, the least common multiple $\mathrm{lcm}(r, s)$. So $\mathrm{ord}(a, b) = \mathrm{lcm}(r, s)$. For $(2, 3)$: $\mathrm{ord}(2) = 6/\gcd(2, 6) = 3$ in $\mathbb{Z}/6\mathbb{Z}$ and $\mathrm{ord}(3) = 8/\gcd(3, 8) = 8$ in $\mathbb{Z}/8\mathbb{Z}$ (Proposition 2.6.4), so $\mathrm{ord}(2, 3) = \mathrm{lcm}(3, 8) = 24$.

(b) ($\Leftarrow$) Assume $\gcd(m,n) = 1$. Let x and y be generators of $\mathbb{Z}/m\mathbb{Z}$ and $\mathbb{Z}/n\mathbb{Z}$ (for instance, the remainders of 1), of orders m and n. By (a), (x,y) has order $\operatorname{lcm}(m,n)$. A common multiple $k = mk'$ of m and n satisfies $n \mid mk'$, hence $n \mid k'$ by Euclid's lemma, so $mn \mid k$; thus $\operatorname{lcm}(m,n) = mn$. So $\mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$, a group of order mn, contains an element of order mn; by Proposition 2.6.1 the cyclic subgroup it generates has mn elements, so it is the whole group. Hence the group is cyclic of order mn, and $\cong \mathbb{Z}/mn\mathbb{Z}$ by Theorem 2.6.2.
($\Rightarrow$) If $d = \gcd(m,n) > 1$, then for any $(a,b) \in \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$, we have $\frac{mn}{d} \cdot (a,b) = (\frac{mn}{d}a, \frac{mn}{d}b) = (0,0)$, since $m \mid \frac{mn}{d}$ and $n \mid \frac{mn}{d}$. So every element has order dividing $\frac{mn}{d} < mn$, and no element has order mn. But $\mathbb{Z}/mn\mathbb{Z}$ has an element of order mn, so the two groups are not isomorphic (Proposition 2.5.9).
(c) Let G have order 4. A non-identity element a cannot have order greater than 4, since then e, a, a^2, a^3, a^4 would be five distinct elements (Proposition 2.6.1). Nor can it have order 3: then $G = \{e, a, a^2, b\}$ for some b, and ab would have to be one of these four elements, but $ab = e$ gives $b = a^{-1} = a^2$, $ab = a$ gives $b = e$, $ab = a^2$ gives $b = a$, and $ab = b$ gives $a = e$. So every non-identity element has order 2 or 4.
Case 1: G has an element a of order 4. Then $G = \{e, a, a^2, a^3\} = \langle a \rangle \cong \mathbb{Z}/4\mathbb{Z}$.
Case 2: Every non-identity element has order 2. Then $G = \{e, a, b, c\}$ with $a^2 = b^2 = c^2 = e$. What is ab? It's not e (else $b = a^{-1} = a$), not a (else $b = e$), and not b (else $a = e$). So $ab = c$, and by the same argument with a and b exchanged, $ba = c$. Likewise $ac = ca = b$ and $bc = cb = a$. So the multiplication table of G is determined, and it matches that of $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ under $e \mapsto (0,0)$, $a \mapsto (1,0)$, $b \mapsto (0,1)$, $c \mapsto (1,1)$. So $G \cong \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$, the ***Klein four-group***.
These two groups are not isomorphic, so up to isomorphism there are exactly two groups of order 4.
(d) Define $\varphi : G \times H \to H \times G$ by $\varphi(g,h) = (h,g)$. It is a bijection, its inverse being $(h,g) \mapsto (g,h)$. It is a homomorphism: $\varphi((g_1,h_1)(g_2,h_2)) = \varphi(g_1g_2, h_1h_2) = (h_1h_2, g_1g_2) = (h_1,g_1)(h_2,g_2) = \varphi(g_1,h_1)\varphi(g_2,h_2)$.

CHAPTER 3

Quotients, Actions, and Presentations

Chapter 2 introduced groups, subgroups, and homomorphisms. In this chapter we learn to divide a group by a subgroup. The cosets of a subgroup $H \leqslant G$ cut G into pieces of equal size, which gives Lagrange's theorem, and when H is normal the cosets themselves form a group G/H. Quotient groups lead to the isomorphism theorems, which describe how quotients, subgroups, and homomorphisms fit together. Along the way we study groups acting on sets, the symmetric group S_n in some detail, and exact sequences; the chapter ends with free groups and presentations, which describe a group by generators and relations.

3.1 Cosets and Quotient Groups

Consider $\mathbb{Z}/n$. Congruence mod n is an equivalence relation on $\mathbb{Z}$, and we may regard $\mathbb{Z}/n$ as its quotient set, whose elements are the classes $a + n\mathbb{Z}$. Addition is compatible with congruence: if $a \equiv a'$ and $b \equiv b' \pmod{n}$, then $a + b \equiv a' + b' \pmod{n}$. So the rule $(a + n\mathbb{Z}) + (b + n\mathbb{Z}) = (a + b) + n\mathbb{Z}$ does not depend on the representatives chosen (this is the criterion of Proposition 1.6.4 for a map to descend to a quotient set). It makes the set of classes a group, and the quotient map $\mathbb{Z} \to \mathbb{Z}/n$, $a \mapsto a + n\mathbb{Z}$, is a homomorphism with kernel $n\mathbb{Z}$.

Why stop at the integers? Given a group G and a subgroup $H \leqslant G$, we can define congruence mod H and form the set of equivalence classes. When does the operation of G descend to these classes, so that the quotient map is a homomorphism? The answer, *normality*, is one of the foundational ideas of group theory.

3.1.1 Congruence Mod H and Cosets

Let $H \leqslant G$. For $a, b \in G$ define $a \sim b$ if and only if $a^{-1}b \in H$. This is an equivalence relation. We have $a \sim a$ since $a^{-1}a = e \in H$. If $a \sim b$, then $a^{-1}b \in H$, so $b^{-1}a = (a^{-1}b)^{-1} \in H$ and $b \sim a$. Finally, if $a \sim b$ and $b \sim c$, then $a^{-1}b$ and $b^{-1}c$ are both in H, so $a^{-1}c = (a^{-1}b)(b^{-1}c) \in H$ since H is closed under the group operation; thus $a \sim c$. We call $\sim$ ***congruence mod*** H.

Definition 3.1.1 (Cosets and index)

Let $H \leqslant G$ and $a \in G$. The ***left coset*** of H containing a is $aH = \{ah \mid h \in H\}$, and the ***right coset*** is $Ha = \{ha \mid h \in H\}$. The set of left cosets is denoted G/H. The ***index*** of H in G, written $[G : H]$, is the number of left cosets of H in G (possibly infinite).

In additive notation the cosets are written $a + H$. For instance, the cosets of $n\mathbb{Z}$ in $\mathbb{Z}$ are the congruence classes $a + n\mathbb{Z}$, and $[\mathbb{Z} : n\mathbb{Z}] = n$.

Proposition 3.1.2 (Cosets partition G)

Let $H \leqslant G$. For $a, b \in G$ we have $b \in aH$ if and only if $a \sim b$. Hence the left cosets of H are exactly the equivalence classes of congruence mod H: they form a partition of G, every $g \in G$ lies in exactly one left coset, namely gH, and $aH = bH$ if and only if $a^{-1}b \in H$.

Proof. We have $b \in aH$ if and only if $b = ah$ for some $h \in H$, that is, if and only if $a^{-1}b \in H$, which is $a \sim b$. So aH is the equivalence class of a. The equivalence classes of an equivalence relation partition the set (Proposition 1.2.8), and two classes are equal exactly when their representatives are equivalent (Lemma 1.2.5). □

Here is the key observation that makes quotient theory work: all cosets have the same size.

Proposition 3.1.3 (Cosets are equinumerous)

For any $a \in G$, the map $\lambda_a : H \to aH$ given by $h \mapsto ah$ is a bijection. In particular, $|aH| = |H|$ for all $a \in G$.

Proof. Surjectivity is immediate from the definition of aH. For injectivity, suppose $ah_1 = ah_2$. Left-multiplying by a^{-1} gives $h_1 = h_2$. (This is left translation, which is always a bijection in a group; we will meet it again in Section 3.2.) □

Theorem 3.1.4 (Lagrange's Theorem)

Let G be a finite group and $H \leqslant G$ a subgroup. Then

$$|G| = [G : H] \cdot |H|.$$

In particular, $|H|$ divides $|G|$, and $[G : H] = |G|/|H|$.

Proof. By Proposition 3.1.2, the $[G : H]$ left cosets of H partition G, and each has $|H|$ elements. Hence $|G| = [G : H] \cdot |H|$. □

Corollary 3.1.5 (Consequences of Lagrange's theorem)

Let G be a finite group.

(a) The order of every element $a \in G$ divides $|G|$.

(b) $a^{|G|} = e$ for every $a \in G$.

(c) If $|G| = p$ is prime, then G is cyclic, generated by any element $a \neq e$, and $G \cong \mathbb{Z}/p$.

Proof. (a) The order of a equals $|\langle a \rangle|$ (Proposition 2.6.1), and $\langle a \rangle$ is a subgroup of G, so its order divides $|G|$ by Lagrange's theorem.

(b) By (a), $|G| = m \cdot \operatorname{ord}(a)$ for some integer m, so $a^{|G|} = (a^{\operatorname{ord}(a)})^m = e$.

(c) Let $a \neq e$. Then $\operatorname{ord}(a)$ divides p and is not 1, so $\operatorname{ord}(a) = p$. Thus $\langle a \rangle$ has $p = |G|$ elements and equals G. A cyclic group of order p is isomorphic to $\mathbb{Z}/p$ (Theorem 2.6.2). □

NOTE The converse of Lagrange's theorem is false: a divisor of $|G|$ need not be the order of any subgroup. The alternating group A_4, of order 12, has no subgroup of order 6; we prove this in Example 3.4.17 once A_4 has been defined. Partial converses do hold. For a finite abelian group every divisor of $|G|$ is the order of a subgroup, as the classification of finite abelian groups shows (Theorem 10.2.1). For an arbitrary finite group, the Sylow theorems of Chapter 11 (Theorem 11.5.3) give, for each prime p, a subgroup whose order is the largest power of p dividing $|G|$.

Example 3.1.6 (Subgroups of S_3)

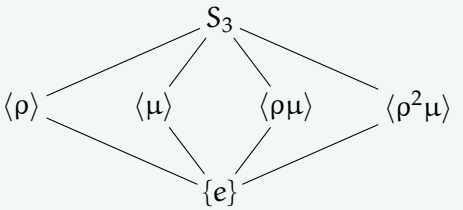
The group S_3 has order 6 and is non-abelian; in particular $S_3 \ncong \mathbb{Z}/6$. Write $\rho = (1\,2\,3)$ and $\mu = (1\,2)$. Composing right to left, $\rho\mu = (1\,3)$ and $\rho^2\mu = (2\,3)$, so

$$S_3 = \{e, \rho, \rho^2, \mu, \rho\mu, \rho^2\mu\} = \{e, (1\,2\,3), (1\,3\,2), (1\,2), (1\,3), (2\,3)\}.$$

By Lagrange's theorem a subgroup of S_3 has order 1, 2, 3, or 6. A subgroup of order 2 or 3 has prime order, so by Corollary 3.1.5 it is generated by an element of order 2 or 3, respectively. The elements of order 2 are the three transpositions, and the elements of order 3 are ρ and $\rho^2 = \rho^{-1}$, which generate the same subgroup. So the subgroups of S_3 are:

- the trivial subgroup $\{e\}$ and S_3 itself;
- one subgroup of order 3, $\langle \rho \rangle = \{e, \rho, \rho^2\} \cong \mathbb{Z}/3$, the rotations of the triangle;
- three subgroups of order 2, $\langle \mu \rangle$, $\langle \rho\mu \rangle$, and $\langle \rho^2\mu \rangle$, each generated by a reflection of the triangle.

In the diagram below (the ***subgroup lattice***), a line joins two subgroups when the lower one is contained in the upper one with no subgroup strictly between them.



Now we ask when G/H is a group. The natural candidate operation is $(aH)(bH) = abH$. For this to make sense it must be well defined: if $aH = a'H$ and $bH = b'H$, we need $abH = a'b'H$. This does *not* always hold.

Example 3.1.7 (Coset multiplication can fail)

Let $G = S_3$ and $H = \{e, (12)\}$, composing right to left (so $(\sigma\tau)(x) = \sigma(\tau(x))$). The three left cosets are $H = \{e, (12)\}$, $(13)H = \{(13), (123)\}$, and $(23)H = \{(23), (132)\}$. Fix $a = (13)$ and take two representatives $b = (23)$ and $b' = (132)$ of the coset $(23)H$. We compute:

$$\begin{aligned} ab &= (13)(23) = (132), \quad \text{so } abH = (132)H = (23)H, \\ ab' &= (13)(132) = (23), \quad \text{so } ab'H = (23)H. \end{aligned}$$

So far the answer does not depend on the representative. Now vary a instead: take $a = (13)$ and $a' = (123)$, which lie in the same coset, with $b = (23)$:

$$\begin{aligned} ab &= (13)(23) = (132), \quad \text{so } abH = (132)H = (23)H, \\ a'b &= (123)(23) = (12), \quad \text{so } a'bH = (12)H = H. \end{aligned}$$

Since $(2\,3)H \neq H$, coset multiplication is not well defined for this H.

What went wrong? The subgroup $H = \{e, (1\,2)\}$ is not "symmetric" enough: its left cosets and right cosets differ. For instance, $(1\,3)H = \{(1\,3), (1\,2\,3)\}$ but $H(1\,3) = \{(1\,3), (1\,3\,2)\}$.

3.1.2 Normal Subgroups and Quotient Groups

For $N \leqslant G$ and $g \in G$, write $gNg^{-1} = \{gng^{-1} \mid n \in N\}$.

Definition 3.1.8 (Normal subgroup)

A subgroup $N \leqslant G$ is ***normal***, written $N \trianglelefteq G$, if $gNg^{-1} = N$ for all $g \in G$; that is, N is closed under conjugation by elements of G.

Proposition 3.1.9 (Equivalent conditions for normality)

For a subgroup $N \leqslant G$, the following are equivalent:

(a) $gNg^{-1} = N$ for all $g \in G$;

(b) $gN = Ng$ for all $g \in G$ (left cosets equal right cosets);

(c) $gNg^{-1} \subseteq N$ for all $g \in G$ (one inclusion suffices).

Proof. (a) $\Rightarrow$ (b): Each element gn of gN equals $(gng^{-1})g$, so $gN = (gNg^{-1})g = Ng$.

(b) $\Rightarrow$ (c): If $gN = Ng$, then for $n \in N$ we have $gn \in gN = Ng$, so $gn = n'g$ for some $n' \in N$, giving $gng^{-1} = n' \in N$.

(c) $\Rightarrow$ (a): We have $gNg^{-1} \subseteq N$ by assumption. For the reverse inclusion, take $n \in N$. Applying (c) to g^{-1} gives $g^{-1}ng \in N$, so $n = g(g^{-1}ng)g^{-1} \in gNg^{-1}$. $\square$

Every subgroup of an abelian group is normal, since there $gng^{-1} = n$. In any group, $\{e\}$ and G are normal.

Theorem 3.1.10 (Quotient groups exist exactly for normal subgroups)

Let K be a subgroup of a group G. The following are equivalent:

(a) $K = \ker \varphi$ for some homomorphism $\varphi : G \to H$ to some group H;

(b) K is normal in G;

(c) the rule $(aK)(bK) = abK$ is a well-defined operation on G/K; that is, $aK = a'K$ and $bK = b'K$ imply $abK = a'b'K$.

When these hold, G/K is a group under this operation, called the ***quotient group*** of G by K, and the map $\pi : G \to G/K$, $\pi(a) = aK$, is a surjective homomorphism with $\ker \pi = K$, called the ***canonical projection*** (or ***quotient map***).

Proof. (a) $\Rightarrow$ (b): If $K = \ker \varphi$, then for $k \in K$ and $g \in G$,

$$\varphi(gkg^{-1}) = \varphi(g)\varphi(k)\varphi(g)^{-1} = \varphi(g)\varphi(g)^{-1} = e,$$

so $gKg^{-1} \subseteq K$ for all g, and K is normal by Proposition 3.1.9.

(b) $\Rightarrow$ (c): Suppose $aK = a'K$ and $bK = b'K$, that is, $a^{-1}a' \in K$ and $b^{-1}b' \in K$. Then

$$(ab)^{-1}(a'b') = b^{-1}a^{-1}a'b' = \big(b^{-1}(a^{-1}a')b\big)\big(b^{-1}b'\big).$$

The first factor lies in $b^{-1}Kb = K$ because K is normal, and the second lies in K. So $(ab)^{-1}(a'b') \in K$, which says $abK = a'b'K$.

Now suppose (c) holds. The group axioms for G/K follow from those of G: $(aK \cdot bK) \cdot cK = (ab)cK = a(bc)K = aK \cdot (bK \cdot cK)$; the coset $eK = K$ is an identity, since $K \cdot aK = aK = aK \cdot K$; and $a^{-1}K$ is inverse to aK. The map π is surjective, and it is a homomorphism because $\pi(ab) = abK = aK \cdot bK = \pi(a)\pi(b)$. Its kernel is $\{a \in G \mid aK = K\} = \{a \in G \mid a \in K\} = K$ by Proposition 3.1.2.

(c) $\Rightarrow$ (a): We have just seen that $K = \ker \pi$. $\square$

In particular, G/H *is a group under* $(aH)(bH) = abH$ *if and only if* H *is normal*, and the normal subgroups of G are exactly the kernels of homomorphisms out of G. The direction (c) $\Rightarrow$ (b) can also be seen directly: if the operation is well defined, then for $h \in H$ and $g \in G$ we have $hH = eH$, so $ghg^{-1}H = (gH)(hH)(g^{-1}H) = (gH)(eH)(g^{-1}H) = H$, that is, $ghg^{-1} \in H$. In additive notation the operation on G/H is $(a + H) + (b + H) = (a + b) + H$; for $H = n\mathbb{Z} \leqslant \mathbb{Z}$ this recovers $\mathbb{Z}/n\mathbb{Z} = \mathbb{Z}/n$.

Proposition 3.1.11 (Subgroups of index two are normal)
If $H \leqslant G$ and $[G : H] = 2$, then $H \trianglelefteq G$.

Proof. If $g \in H$, then $gH = H = Hg$. Let $g \notin H$. The two left cosets are H and gH, and they partition G, so $gH = G \setminus H$. The right coset Hg is disjoint from H: if $hg = h'$ with $h, h' \in H$, then $g = h^{-1}h' \in H$. Hence $Hg \subseteq G \setminus H = gH$. The same argument applied to $g^{-1} \notin H$ gives $Hg^{-1} \subseteq g^{-1}H$, and taking inverses of all elements turns this into $gH \subseteq Hg$. So $gH = Hg$. $\square$

Example 3.1.12 (Normal and non-normal subgroups of S_3)

In S_3, the subgroup $A_3 = \langle (1\,2\,3) \rangle = \{e, (1\,2\,3), (1\,3\,2)\}$ has index 2, so it is normal and S_3/A_3 is a group of order 2. The subgroup $H = \{e, (1\,2)\}$ is not normal: $(1\,2\,3)(1\,2)(1\,2\,3)^{-1} = (2\,3)$, so $(1\,2\,3)H(1\,2\,3)^{-1} = \{e, (2\,3)\} \neq H$. Similarly, conjugation by $(1\,2\,3)$ carries $\langle (1\,3) \rangle$ to $\langle (1\,2) \rangle$ and $\langle (2\,3) \rangle$ to $\langle (1\,3) \rangle$, so none of the three subgroups of order 2 is normal. Hence the normal subgroups of S_3 are $\{e\}$, A_3, and S_3.

Definition 3.1.13 (Center)

The ***center*** of a group G is $Z(G) = \{z \in G \mid zg = gz \text{ for all } g \in G\}$.

Proposition 3.1.14 (The center)

$Z(G)$ is an abelian normal subgroup of G, and G is abelian if and only if $Z(G) = G$.

Proof. Clearly $e \in Z(G)$. If $z, z' \in Z(G)$, then $zz'g = zgz' = gzz'$ for all g, so $zz' \in Z(G)$; multiplying $zg = gz$ by z^{-1} on both sides gives $gz^{-1} = z^{-1}g$, so $z^{-1} \in Z(G)$. Elements of $Z(G)$ commute with everything, in particular with each other, so $Z(G)$ is abelian. For $z \in Z(G)$ and $g \in G$, $gzg^{-1} = zgg^{-1} = z$, so $gZ(G)g^{-1} \subseteq Z(G)$ and $Z(G)$ is normal. Finally, $Z(G) = G$ says exactly that all elements commute. □

Question 3.1.15

Let G be a group.

(a) Prove that if $G/Z(G)$ is cyclic, then G is abelian.

(b) Find $Z(D_4)$ and describe $D_4/Z(D_4)$.

Solution. **(a)** Write $Z = Z(G)$ and suppose $G/Z = \langle aZ \rangle$. Every coset is then $(aZ)^k = a^kZ$ for some $k \in \mathbb{Z}$, so every element of G has the form a^kz with $k \in \mathbb{Z}$ and $z \in Z$. Take $g = a^iz_1$ and $h = a^jz_2$. Since z_1 and z_2 commute with everything,

$$gh = a^iz_1a^jz_2 = a^{i+j}z_1z_2 \quad \text{and} \quad hg = a^jz_2a^iz_1 = a^{i+j}z_2z_1 = a^{i+j}z_1z_2.$$

Thus $gh = hg$, and G is abelian. (So in fact $G/Z(G)$ is cyclic only when it is trivial.)

(b) Recall (Proposition 2.2.15) that $D_4 = \{e, r, r^2, r^3, s, sr, sr^2, sr^3\}$ with $r^4 = s^2 = e$ and $srs^{-1} = r^{-1}$. From $sr = r^{-1}s$ we get $sr^k = r^{-k}s$ for all k by induction. An element z is central if and only if it commutes with the generators r and s, because the elements commuting with z form a subgroup.
For rotations: r^k commutes with s if and only if $r^{-k}s = r^k s$, that is, $r^{2k} = e$, so $k \in \{0, 2\}$. Thus e and r^2 are the central rotations.
For reflections: sr^j commutes with r if and only if $r \cdot sr^j = sr^j \cdot r$. Now $rsr^j = r \cdot r^{-j}s = r^{1-j}s$, while $sr^{j+1} = r^{-j-1}s$. Equality would give $r^{1-j} = r^{-j-1}$, so $r^2 = e$, which is false since r has order 4. No reflection is central.
Therefore $Z(D_4) = \{e, r^2\} \cong \mathbb{Z}/2$. The quotient $D_4/Z(D_4)$ has order $8/2 = 4$, with cosets $\{e, r^2\}, \{r, r^3\}, \{s, sr^2\}, \{sr, sr^3\}$. Every non-identity element of the quotient has order 2 (since r^2, s^2, and $(sr)^2$ lie in $Z(D_4)$), so $D_4/Z(D_4) \cong \mathbb{Z}/2 \times \mathbb{Z}/2$, the Klein four-group, and not $\mathbb{Z}/4$.

Normal subgroups let us view a group G as built from two pieces, a normal subgroup N and the quotient G/N. The groups that cannot be broken up this way are the basic building blocks.

Definition 3.1.16 (Simple groups)

A group $G \neq \{e\}$ is ***simple*** if its only normal subgroups are $\{e\}$ and G.

By Lagrange's theorem, a group of prime order has no subgroups other than $\{e\}$ and itself, so $\mathbb{Z}/p$ is simple for every prime p. The alternating groups A_n with $n \geqslant 5$ (Section 3.4) are simple as well; this is proved in Chapter 11 (Theorem 11.4.9).

Question 3.1.17

Let G be a group of order $2p$, where p is an odd prime.

(a) Prove that G has a unique subgroup of order p, which is therefore normal.

(b) Conclude that $G \cong \mathbb{Z}/2p$ or $G \cong D_p$, the dihedral group of order $2p$.

Solution. **(a)** *Existence.* By Corollary 3.1.5, the order of every element divides $2p$, so it is $1, 2, p$, or $2p$. If some g has order $2p$, then g^2 has order p. Suppose no element has order p or $2p$. Then $g^2 = e$ for every $g \in G$, which forces G to be abelian: $ab = (ab)^{-1} = b^{-1}a^{-1} = ba$. Choose distinct non-identity elements a, b (possible since $|G| \geqslant 6$). Then $\{e, a, b, ab\}$ is closed under

multiplication and inverses, so it is a subgroup of order 4. But 4 does not divide $2p$, contradicting Lagrange's theorem. Hence some $a \in G$ has order p, and $H = \langle a \rangle$ is a subgroup of order p. It has index 2, so it is normal by Proposition 3.1.11.

Uniqueness. Suppose K is a subgroup of order p with $K \neq H$. Then $H \cap K$ is a subgroup of H (Proposition 2.4.4) whose order divides p and is not p, so $H \cap K = \{e\}$. The map $H \times K \to G$, $(h, k) \mapsto hk$, is injective: if $hk = h'k'$, then $h'^{-1}h = k'k^{-1} \in H \cap K = \{e\}$, so $h = h'$ and $k = k'$. Hence $|G| \geqslant |H||K| = p^2 > 2p$, a contradiction.

(b) Let $H = \langle a \rangle$ be the unique subgroup of order p, and pick $b \in G \setminus H$. The order of b is not 1, and it is not p, for then $\langle b \rangle$ would be a second subgroup of order p. If b has order $2p$, then $G = \langle b \rangle$ is cyclic and $G \cong \mathbb{Z}/2p$. It remains to treat the case $\operatorname{ord}(b) = 2$.

Since H is normal, $bab^{-1} \in H$, say $bab^{-1} = a^k$. Conjugation by b is a homomorphism, since $(bxb^{-1})(byb^{-1}) = bxyb^{-1}$; using this and $b^2 = e$,

$$a = b^2ab^{-2} = b(bab^{-1})b^{-1} = ba^kb^{-1} = (bab^{-1})^k = a^{k^2},$$

so p divides $k^2 - 1 = (k-1)(k+1)$, and by Euclid's lemma $k \equiv \pm 1 \pmod{p}$. As $b \notin H$, the two cosets of H are H and bH, so every element of G is uniquely a^i or a^ib with $0 \leqslant i < p$.

If $k \equiv 1$, then $ab = ba$, so G is abelian. Then $(ab)^m = a^mb^m$ equals e if and only if $a^m = b^{-m} \in H \cap \langle b \rangle = \{e\}$, that is, if and only if both p and 2 divide m. So ab has order $2p$, and $G \cong \mathbb{Z}/2p$.

If $k \equiv -1$, then $bab^{-1} = a^{-1}$, so $ba^j = a^{-j}b$ for all j, and

$$(a^ib^\varepsilon)(a^jb^\delta) = a^{i+(-1)^\varepsilon j}\, b^{\varepsilon+\delta} \qquad (\varepsilon, \delta \in \{0, 1\}).$$

The dihedral group D_p consists of the elements r^is^ε ($0 \leqslant i < p$, $\varepsilon \in \{0, 1\}$) with $r^p = s^2 = e$ and $srs^{-1} = r^{-1}$ (Proposition 2.2.15), so its elements multiply by the same rule. Hence $a^ib^\varepsilon \mapsto r^is^\varepsilon$ is a bijective homomorphism, and $G \cong D_p$.

3.2 Group Actions

The notion of a group "acting" on a set underlies much of mathematics. When we study symmetries of geometric objects, permutations of the roots of a polynomial, or even the structure of a group itself, we are studying group actions. Understanding which sets a group G acts on, and how, gives information about G. These ideas return throughout the book: in Chapter 11 actions are the main tool

for analyzing finite groups, and in representation theory groups act on vector spaces by linear maps.

3.2.1 Actions, Orbits, and Stabilizers

Definition 3.2.1 (Group action)

A ***(left) action*** of a group G on a set X is a map $G \times X \to X$, written $(g, x) \mapsto g \cdot x$, satisfying:

(a) $e \cdot x = x$ for all $x \in X$;

(b) $(gh) \cdot x = g \cdot (h \cdot x)$ for all $g, h \in G$ and $x \in X$.

We say G ***acts on*** X, or that X is a G***-set***.

Recall (Definition 2.2.8) that $\mathrm{Sym}(X)$ denotes the group of all bijections $X \to X$ under composition.

Proposition 3.2.2 (Actions are homomorphisms to $\mathrm{Sym}(X)$)

Given an action of G on X, each map $\rho(g) : X \to X$, $x \mapsto g \cdot x$, is a bijection, and $\rho : G \to \mathrm{Sym}(X)$ is a homomorphism. Conversely, every homomorphism $\rho : G \to \mathrm{Sym}(X)$ defines an action by $g \cdot x = \rho(g)(x)$.

Proof. By the two axioms, $\rho(g)(\rho(g^{-1})(x)) = (gg^{-1}) \cdot x = e \cdot x = x$, and likewise $\rho(g^{-1})(\rho(g)(x)) = x$. So $\rho(g)$ is a bijection with inverse $\rho(g^{-1})$. The second axiom says $\rho(gh) = \rho(g) \circ \rho(h)$, so ρ is a homomorphism. Conversely, if ρ is a homomorphism, then $\rho(e) = \mathrm{id}_X$ gives $e \cdot x = x$, and $\rho(gh) = \rho(g) \circ \rho(h)$ gives $(gh) \cdot x = g \cdot (h \cdot x)$. $\square$

So a group action is the same thing as a homomorphism from G to a group of permutations. In Chapter 13 we replace $\mathrm{Sym}(X)$ by the group of invertible linear maps of a vector space; such homomorphisms are the *representations* of G (Definition 13.1.1).

Definition 3.2.3 (Orbits and stabilizers)

Let G act on X and let $x \in X$.

(a) The ***orbit*** of x is $G \cdot x = \{g \cdot x \mid g \in G\} \subseteq X$, also written $\mathcal{O}_x$.

(b) The ***stabilizer*** of x is $\mathrm{Stab}(x) = \{g \in G \mid g \cdot x = x\} \subseteq G$.

Proposition 3.2.4 (Stabilizers are subgroups)

For any $x \in X$, the stabilizer $\mathrm{Stab}(x)$ is a subgroup of G.

Proof. We have $e \in \mathrm{Stab}(x)$ since $e \cdot x = x$. If $g, h \in \mathrm{Stab}(x)$, then $(gh) \cdot x = g \cdot (h \cdot x) = g \cdot x = x$, so $gh \in \mathrm{Stab}(x)$. If $g \in \mathrm{Stab}(x)$, then $x = e \cdot x = (g^{-1}g) \cdot x = g^{-1} \cdot (g \cdot x) = g^{-1} \cdot x$, so $g^{-1} \in \mathrm{Stab}(x)$. □

Proposition 3.2.5 (Orbits partition X)

Define $x \sim y$ if $y = g \cdot x$ for some $g \in G$. This is an equivalence relation on X, and its equivalence classes are the orbits. Hence the orbits form a partition of X.

Proof. Reflexivity: $x = e \cdot x$. Symmetry: if $y = g \cdot x$, then $g^{-1} \cdot y = (g^{-1}g) \cdot x = x$. Transitivity: if $y = g \cdot x$ and $z = h \cdot y$, then $z = h \cdot (g \cdot x) = (hg) \cdot x$. The class of x is $\{g \cdot x \mid g \in G\} = G \cdot x$. □

Definition 3.2.6 (Faithful and transitive actions)

Let G act on X, with associated homomorphism $\rho : G \to \mathrm{Sym}(X)$.

(a) The ***kernel*** of the action is $\ker \rho = \{g \in G \mid g \cdot x = x \text{ for all } x \in X\}$. The action is ***faithful*** if ρ is injective, that is, if only e acts as the identity on X.

(b) The action is ***transitive*** if $X \neq \emptyset$ and there is only one orbit: for all $x, y \in X$ there exists $g \in G$ with $g \cdot x = y$.

If an action is not faithful, the group that "really" acts on X is $G/\ker \rho$. Any action of G on X restricts to an action on each orbit, and that action is transitive; so every action is a disjoint union of transitive actions.

Example 3.2.7 (Fundamental examples of group actions)

(a) **Permutations:** S_n acts on $\{1, \ldots, n\}$ by $\sigma \cdot i = \sigma(i)$. This action is faithful (the associated homomorphism is the identity of S_n) and transitive (the transposition $(i\,j)$ sends i to j).

(b) **Left multiplication:** G acts on itself by $g \cdot h = gh$. This action is faithful, since $g \cdot e = g$, and transitive, since $(yx^{-1}) \cdot x = y$. Every stabilizer is trivial.

(c) **The coset action:** If $H \leqslant G$, then G acts on the set G/H of left cosets by $g \cdot (aH) = (ga)H$. This is well defined: if $aH = a'H$, then $a' = ah$ for some $h \in H$, so $ga'H = gahH = gaH$. The action is transitive, since $(ba^{-1}) \cdot aH = bH$. The stabilizer of aH is aHa^{-1}, because $gaH = aH$ if and only if $a^{-1}ga \in H$.

(d) **Conjugation:** G acts on itself by $g \cdot h = ghg^{-1}$. This action is the subject of Section 3.2.3 below.

(e) **Conjugation on subgroups:** G acts on the set of its subgroups by $g \cdot K = gKg^{-1}$. (The set gKg^{-1} is a subgroup, being the image of K under the homomorphism $h \mapsto ghg^{-1}$ of Section 3.2.3.) The subgroups fixed by every $g \in G$ are exactly the normal subgroups.

(f) **Linear actions:** $GL_n(\mathbb{R})$ acts on $\mathbb{R}^n$ by matrix multiplication, $A \cdot v = Av$. This is the prototype for representation theory.

Proposition 3.2.8 (Stabilizers along an orbit are conjugate)
If $y = g \cdot x$, then $\mathrm{Stab}(y) = g\,\mathrm{Stab}(x)g^{-1}$.

Proof. If $h \in \mathrm{Stab}(x)$, then $(ghg^{-1}) \cdot y = (ghg^{-1}g) \cdot x = g \cdot (h \cdot x) = g \cdot x = y$, so $g\,\mathrm{Stab}(x)g^{-1} \subseteq \mathrm{Stab}(y)$. Since $x = g^{-1} \cdot y$, the same argument gives $g^{-1}\,\mathrm{Stab}(y)g \subseteq \mathrm{Stab}(x)$, that is, $\mathrm{Stab}(y) \subseteq g\,\mathrm{Stab}(x)g^{-1}$. □

3.2.2 The Orbit–Stabilizer Theorem

The following theorem is one of the most useful results in group theory. It says that the action of G on an orbit looks exactly like the coset action of Example 3.2.7(c), and it connects the size of an orbit to the index of a stabilizer.

Theorem 3.2.9 (Orbit–Stabilizer Theorem)
Let G act on X, let $x \in X$, and let $H = \mathrm{Stab}(x)$. The map

$$\varepsilon : G/H \to G \cdot x, \qquad aH \mapsto a \cdot x,$$

is a well-defined bijection, and it is ***equivariant***: $\varepsilon(g \cdot aH) = g \cdot \varepsilon(aH)$ for all $g, a \in G$. In particular $|G \cdot x| = [G : \mathrm{Stab}(x)]$, and if G is finite,

$$|G \cdot x| = \frac{|G|}{|\mathrm{Stab}(x)|},$$

so the size of every orbit divides $|G|$.

Proof. **Well defined:** If $aH = bH$, then $b = ah$ for some $h \in H$, so $b \cdot x = a \cdot (h \cdot x) = a \cdot x$.

Injective: If $a \cdot x = b \cdot x$, then $(a^{-1}b) \cdot x = a^{-1} \cdot (b \cdot x) = a^{-1} \cdot (a \cdot x) = x$, so $a^{-1}b \in H$ and $aH = bH$.

Surjective: Every element of $G \cdot x$ has the form $a \cdot x = \varepsilon(aH)$.

Equivariant: $\varepsilon(g \cdot aH) = \varepsilon(gaH) = (ga) \cdot x = g \cdot (a \cdot x) = g \cdot \varepsilon(aH)$.

The counting statements follow from the bijection and Lagrange's theorem. □

Corollary 3.2.10 (Counting orbits)

Let a finite group G act on a finite set X, and let $x_1, \dots, x_m$ be one element from each orbit. Then

$$|X| = \sum_{i=1}^{m} |G \cdot x_i| = \sum_{i=1}^{m} [G : \mathrm{Stab}(x_i)].$$

Proof. The orbits partition X (Proposition 3.2.5); now apply the Orbit–Stabilizer Theorem to each orbit. □

Example 3.2.11

Consider S_4 acting on $\{1,2,3,4\}$ in the natural way, and take $x = 1$. The orbit is $S_4 \cdot 1 = \{1,2,3,4\}$, since the action is transitive. The stabilizer $\mathrm{Stab}(1) = \{\sigma \in S_4 \mid \sigma(1) = 1\}$ consists of the permutations of $\{2,3,4\}$, a group isomorphic to S_3, so $|\mathrm{Stab}(1)| = 6$. Check: $|S_4 \cdot 1| = 4 = 24/6 = |S_4|/|\mathrm{Stab}(1)|$.

3.2.3 The Conjugation Action and Conjugacy Classes

A group G acts on itself by conjugation, $g \cdot h = ghg^{-1}$. The axioms hold: $e \cdot h = h$, and

$$(gg') \cdot h = gg'h(gg')^{-1} = g(g'hg'^{-1})g^{-1} = g \cdot (g' \cdot h).$$

Write $c_g(h) = ghg^{-1}$. Each c_g is a homomorphism, since $c_g(hk) = ghkg^{-1} = (ghg^{-1})(gkg^{-1})$, and a bijection with inverse $c_{g^{-1}}$. So c_g is an automorphism of G, and by Proposition 3.2.2 the action is a homomorphism

$$c : G \to \mathrm{Aut}(G), \qquad g \mapsto c_g.$$

The automorphisms c_g are called ***inner automorphisms***, and their set $\mathrm{Inn}(G) = \mathrm{Im}\, c$ is a subgroup of $\mathrm{Aut}(G)$.

Definition 3.2.12 (Conjugacy classes and centralizers)

Let G be a group. Two elements $h, h' \in G$ are ***conjugate*** if $h' = ghg^{-1}$ for some $g \in G$. The ***conjugacy class*** of h is its orbit under conjugation,

$$C_h = \{ghg^{-1} \mid g \in G\},$$

and the ***centralizer*** of h is its stabilizer, $Z(h) = \{g \in G \mid gh = hg\}$, the subgroup of elements commuting with h.

Proposition 3.2.13 (Properties of conjugacy classes)

Let G be a group.

(a) The conjugacy classes of G form a partition of G.

(b) $|C_h| = [G : Z(h)]$; if G is finite, $|C_h|$ divides $|G|$.

(c) $C_h = \{h\}$ if and only if $h \in Z(G)$. The kernel of the conjugation action is $Z(G)$, so the action is faithful if and only if $Z(G) = \{e\}$.

(d) A subgroup $N \leqslant G$ is normal if and only if it is a union of conjugacy classes.

Proof. (a) is Proposition 3.2.5, and (b) is the Orbit–Stabilizer Theorem, since $Z(h) = \mathrm{Stab}(h)$.

(c) $C_h = \{h\}$ means $ghg^{-1} = h$, that is, $gh = hg$, for all g; this says $h \in Z(G)$. An element g is in the kernel when $ghg^{-1} = h$ for all h, that is, when $g \in Z(G)$.

(d) If N is normal and $n \in N$, then $C_n = \{gng^{-1}\} \subseteq N$, so N is the union of the classes of its elements. Conversely, if N is a union of conjugacy classes, then $gng^{-1} \in C_n \subseteq N$ for all $n \in N$ and $g \in G$, so N is normal by Proposition 3.1.9. □

The conjugacy classes of S_n are described in Section 3.4, and Chapter 11 builds on this action to prove the class equation and the Sylow theorems.

3.2.4 Cayley's Theorem

Theorem 3.2.14 (Cayley's Theorem)

Every group G is isomorphic to a subgroup of $\mathrm{Sym}(G)$. If $|G| = n$ is finite, then G is isomorphic to a subgroup of S_n.

Proof. The left multiplication action of G on itself (Example 3.2.7(b)) gives a homomorphism $\rho : G \to \mathrm{Sym}(G)$, $\rho(g)(h) = gh$. The action is faithful: if $\rho(g) = \mathrm{id}_G$, then $g = g \cdot e = e$. So $\ker \rho = \{e\}$, ρ is injective, and $G \cong \mathrm{Im}\,\rho \leqslant \mathrm{Sym}(G)$ (Proposition 2.5.6).

If $|G| = n$, choose a bijection $\beta : G \to \{1, \dots, n\}$. Then $\sigma \mapsto \beta\sigma\beta^{-1}$ is an isomorphism $\mathrm{Sym}(G) \to S_n$, with inverse $\tau \mapsto \beta^{-1}\tau\beta$, since $\beta\sigma\sigma'\beta^{-1} = (\beta\sigma\beta^{-1})(\beta\sigma'\beta^{-1})$. Composing gives an injective homomorphism $G \to S_n$. □

NOTE Cayley's theorem is simultaneously profound and useless. Profound: every abstract group is "really" a group of permutations, so the group axioms capture exactly the structure of composition of bijections. Useless: the embedding is almost always far larger than necessary. S_n has order $n!$, so embedding a group of order n into S_n wastes enormous space. Better embeddings come from smaller faithful actions.

The coset action gives such smaller actions. For $H \leqslant G$, the intersection $\bigcap_{a \in G} aHa^{-1}$ is called the ***core*** of H in G.

Proposition 3.2.15 (Actions on cosets)

Let $H \leqslant G$ with $[G : H] = n$ finite. The action of G on G/H gives a homomorphism $\rho : G \to \mathrm{Sym}(G/H) \cong S_n$ whose kernel is the core $\bigcap_{a \in G} aHa^{-1}$. The core is the largest normal subgroup of G contained in H: it is normal, it is contained in H, and it contains every normal subgroup of G contained in H. In particular, if H contains no normal subgroup of G other than $\{e\}$, then G is isomorphic to a subgroup of S_n.

Proof. The kernel of the action is the set of elements fixing every coset, which is $\bigcap_{a \in G} \mathrm{Stab}(aH) = \bigcap_{a \in G} aHa^{-1}$ by Example 3.2.7(c). As a kernel it is normal (Theorem 3.1.10), and taking $a = e$ shows it is contained in H. If $K \trianglelefteq G$ and $K \subseteq H$, then $K = aKa^{-1} \subseteq aHa^{-1}$ for every $a \in G$, so K is contained in the core. For the last statement: the kernel is then $\{e\}$, so ρ is injective and G is isomorphic to its image (Proposition 2.5.6), a subgroup of $\mathrm{Sym}(G/H) \cong S_n$. □

Example 3.2.16

(a) Let $G = S_4$ and $H = \mathrm{Stab}(4) \cong S_3$. Then $[G : H] = 4$, and by the Orbit–Stabilizer Theorem the action on G/H is the natural action of S_4 on $\{1, 2, 3, 4\}$; here ρ is an isomorphism.

(b) Let $G = D_4$, of order 8, and $H = \{e, s\}$ for a reflection s, so $[G : H] = 4$. Cayley's theorem embeds D_4 in S_8, but we can do better. From $srs^{-1} = r^{-1}$ we get $sr^{-1} = rs$, so $rsr^{-1} = r(sr^{-1}) = r^2 s \neq s$. Thus $H \cap rHr^{-1} = \{e, s\} \cap \{e, r^2 s\} = \{e\}$, the core of H is trivial, and D_4 is isomorphic to a subgroup of S_4. (If s fixes a vertex of the square, then

H is the stabilizer of that vertex, and by the Orbit–Stabilizer Theorem this is the action of D_4 on the four vertices.)

3.3 The Isomorphism Theorems

The isomorphism theorems describe how quotients, subgroups, and homomorphisms interact. They reappear essentially unchanged for rings, vector spaces, and modules; Chapter 8 develops a language in which this common pattern can be stated once.

3.3.1 The First Isomorphism Theorem

Recall the kernel and image of a homomorphism $\varphi : G \to H$ (Definition 2.5.5): $\ker \varphi$ is a subgroup of G, $\operatorname{Im} \varphi$ is a subgroup of H, and φ is injective if and only if $\ker \varphi = \{e\}$ (Proposition 2.5.6). By Theorem 3.1.10, $\ker \varphi$ is moreover a normal subgroup, so $G/\ker \varphi$ is a group.

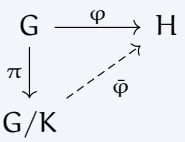
Lemma 3.3.1 (Fibers of a homomorphism)
Let $\varphi : G \to H$ be a homomorphism and $a, b \in G$. Then

$$\varphi(a) = \varphi(b) \iff a^{-1}b \in \ker \varphi \iff a \ker \varphi = b \ker \varphi.$$

Proof. $\varphi(a) = \varphi(b) \iff \varphi(a)^{-1}\varphi(b) = e \iff \varphi(a^{-1}b) = e \iff a^{-1}b \in \ker \varphi$. The last equivalence is Proposition 3.1.2. □

When does a homomorphism out of G induce a homomorphism out of a quotient G/K? Exactly when it kills K.

Theorem 3.3.2 (Factoring through a quotient)
Let $\varphi : G \to H$ be a homomorphism, let $K \trianglelefteq G$, and let $\pi : G \to G/K$ be the canonical projection. There exists a map $\bar{\varphi} : G/K \to H$ with $\varphi = \bar{\varphi} \circ \pi$ if and only if $K \subseteq \ker \varphi$.



In that case $\bar{\varphi}$ is unique, given by $\bar{\varphi}(aK) = \varphi(a)$; it is a homomorphism; $\operatorname{Im} \bar{\varphi} = \operatorname{Im} \varphi$; and $\ker \bar{\varphi} = \pi(\ker \varphi) = (\ker \varphi)/K$.

Proof. If $\bar{\varphi}$ exists, then for $k \in K$ we have $\varphi(k) = \bar{\varphi}(kK) = \bar{\varphi}(eK) = \varphi(e) = e$, so $K \subseteq \ker \varphi$.

Conversely, suppose $K \subseteq \ker \varphi$. If $aK = bK$, then $a^{-1}b \in K \subseteq \ker \varphi$, so $\varphi(a) = \varphi(b)$ by Lemma 3.3.1. Thus φ is constant on each coset of K, and $\bar{\varphi}(aK) = \varphi(a)$ is well defined (Proposition 1.6.4); by construction $\varphi = \bar{\varphi} \circ \pi$. Any $\bar{\varphi}$ with $\varphi = \bar{\varphi} \circ \pi$ must be given by this formula, because π is surjective. It is a homomorphism:

$$\bar{\varphi}(aK \cdot bK) = \bar{\varphi}(abK) = \varphi(ab) = \varphi(a)\varphi(b) = \bar{\varphi}(aK)\,\bar{\varphi}(bK).$$

Its image is $\{\varphi(a) \mid a \in G\} = \operatorname{Im} \varphi$. Finally, $\bar{\varphi}(aK) = e$ if and only if $a \in \ker \varphi$, so $\ker \bar{\varphi} = \{aK \mid a \in \ker \varphi\} = \pi(\ker \varphi)$. □

Theorem 3.3.3 (First Isomorphism Theorem)

Let $\varphi : G \to H$ be a homomorphism. Then $\ker \varphi \trianglelefteq G$, and the map

$$\bar{\varphi} : G/\ker \varphi \to \operatorname{Im} \varphi, \qquad a \ker \varphi \mapsto \varphi(a),$$

is an isomorphism. In particular, if φ is surjective, then $G/\ker \varphi \cong H$.

Proof. Normality of $\ker \varphi$ is Theorem 3.1.10. Apply Theorem 3.3.2 with $K = \ker \varphi$: the map $\bar{\varphi}$ is a homomorphism with image $\operatorname{Im} \varphi$ and kernel $\pi(\ker \varphi) = \{\ker \varphi\}$, the identity element of $G/\ker \varphi$. So $\bar{\varphi}$ is injective (Proposition 2.5.6) and surjective onto $\operatorname{Im} \varphi$. □

In slogan form: *the image of a homomorphism is isomorphic to the domain modulo the kernel.*

NOTE The First Isomorphism Theorem is sometimes called the "fundamental theorem of homomorphisms." It says that every homomorphism $\varphi : G \to H$ factors as

$$G \twoheadrightarrow G/\ker \varphi \xrightarrow{\sim} \operatorname{Im} \varphi \hookrightarrow H,$$

a surjection, followed by an isomorphism, followed by an inclusion.

Example 3.3.4 (The First Isomorphism Theorem at work)

(a) Reduction mod n (Example 2.5.7) is a surjective homomorphism from $\mathbb{Z}$ onto the group $\{0, 1, \ldots, n-1\}$ with addition mod n of Example 2.2.3, with kernel $n\mathbb{Z}$. So the quotient group $\mathbb{Z}/n\mathbb{Z}$ is isomorphic to that group, and the two descriptions of $\mathbb{Z}/n$ agree.

(b) The map $t \mapsto e^{2\pi i t}$ is a surjective homomorphism from $(\mathbb{R}, +)$ onto the circle group $S^1 = \{z \in \mathbb{C} \mid |z| = 1\}$ under multiplication, with kernel $\mathbb{Z}$. Hence $\mathbb{R}/\mathbb{Z} \cong S^1$. Each coset $t + \mathbb{Z}$ contains exactly one number in $[0, 1)$, namely the fractional part of t, so $\mathbb{R}/\mathbb{Z}$ is the group $[0, 1)$ with addition mod 1 of Example 2.2.4.

(c) The determinant (Section 5.4) is a homomorphism $\det : \mathrm{GL}_n(\mathbb{R}) \to \mathbb{R}^\times$. It is surjective, since $\det \mathrm{diag}(t, 1, \ldots, 1) = t$, and its kernel is $\mathrm{SL}_n(\mathbb{R})$. Hence $\mathrm{GL}_n(\mathbb{R})/\mathrm{SL}_n(\mathbb{R}) \cong \mathbb{R}^\times$.

(d) The conjugation homomorphism $c : G \to \mathrm{Aut}(G)$ of Section 3.2.3 has image $\mathrm{Inn}(G)$ and kernel $Z(G)$ (Proposition 3.2.13). Hence $G/Z(G) \cong \mathrm{Inn}(G)$.

3.3.2 The Second and Third Isomorphism Theorems

For subsets $A, B \subseteq G$ we write $AB = \{ab \mid a \in A, b \in B\}$.

Theorem 3.3.5 (Second Isomorphism Theorem)

Let G be a group, $H \leqslant G$ a subgroup, and $N \trianglelefteq G$ a normal subgroup. Then:

(a) HN is a subgroup of G, and $N \trianglelefteq HN$;

(b) $H \cap N \trianglelefteq H$;

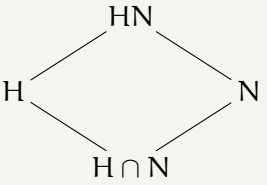
(c) $H/(H \cap N) \cong HN/N$, via $h(H \cap N) \mapsto hN$.

Proof. **(a)** We have $e = e \cdot e \in HN$. For closure, $(h_1 n_1)(h_2 n_2) = h_1 h_2 (h_2^{-1} n_1 h_2) n_2 \in HN$, since $h_2^{-1} n_1 h_2 \in N$ by normality. For inverses, $(hn)^{-1} = n^{-1} h^{-1} = h^{-1}(h n^{-1} h^{-1}) \in HN$, since $h n^{-1} h^{-1} \in N$. Since N is closed under conjugation by all of G, it is in particular normal in HN.

(b) $H \cap N$ is a subgroup of H (Proposition 2.4.4). If $x \in H \cap N$ and $h \in H$, then $hxh^{-1} \in H$ because H is a subgroup, and $hxh^{-1} \in N$ because $N \trianglelefteq G$; so $hxh^{-1} \in H \cap N$.

(c) Define $\varphi : H \to HN/N$ by $\varphi(h) = hN$. This is a homomorphism, being the restriction to H of the canonical projection $HN \to HN/N$. It is surjective: every element of HN/N has the form $hnN = hN = \varphi(h)$. Its kernel is $\{h \in H \mid hN = N\} = \{h \in H \mid h \in N\} = H \cap N$. By the First Isomorphism Theorem, $H/(H \cap N) \cong HN/N$. □

NOTE The name "Diamond Theorem" sometimes used for the Second Isomorphism Theorem comes from the picture of the four subgroups involved:



The theorem identifies the quotient along the lower-left edge, $H/(H \cap N)$, with the quotient along the opposite, upper-right edge, HN/N. (The other two edges are not paired in general: $N/(H \cap N)$ need not even be a group, since $H \cap N$ need not be normal in N.) For finite groups, comparing orders gives $|HN| = |H|\,|N|/|H \cap N|$.

Theorem 3.3.6 (Third Isomorphism Theorem)

Let G be a group and let $N \subseteq M$ be normal subgroups of G. Then:

(a) M/N is a normal subgroup of G/N;

(b) $(G/N)/(M/N) \cong G/M$, via $gN(M/N) \mapsto gM$.

Proof. Define $\psi : G/N \to G/M$ by $\psi(gN) = gM$. This is well defined by Theorem 3.3.2 applied to the projection $G \to G/M$, whose kernel M contains N; and it is a surjective homomorphism, since $G \to G/M$ is. Its kernel is $\{gN \mid gM = M\} = \{gN \mid g \in M\} = M/N$. So M/N is a kernel, hence normal in G/N, and by the First Isomorphism Theorem $(G/N)/(M/N) \cong \operatorname{Im}\psi = G/M$. □

NOTE The Third Isomorphism Theorem says "quotients of quotients are quotients": $(G/N)/(M/N) \cong G/M$, as if the N's cancelled. This is exactly what one would hope for, and it is a sign that quotient groups are a natural and well-behaved construction.

3.3.3 The Correspondence Theorem

The last structural theorem describes the subgroups of a quotient in terms of the subgroups of the original group.

Theorem 3.3.7 (Correspondence Theorem)

Let $N \trianglelefteq G$ and let $\pi : G \to G/N$ be the canonical projection. There is a bijection

$$\{\text{subgroups } H \text{ of } G \text{ with } N \subseteq H\} \longleftrightarrow \{\text{subgroups of } G/N\}$$

given by $H \mapsto H/N = \pi(H)$, with inverse $\bar{H} \mapsto \pi^{-1}(\bar{H})$. Moreover, for subgroups H, H_1, H_2 containing N:

(a) $H_1 \subseteq H_2$ if and only if $H_1/N \subseteq H_2/N$, and in that case $[H_2 : H_1] = [H_2/N : H_1/N]$;

(b) $H \trianglelefteq G$ if and only if $H/N \trianglelefteq G/N$.

Proof. The image of a subgroup under a homomorphism is a subgroup, and so is the preimage (Proposition 2.5.2). So $\pi(H)$ is a subgroup of G/N, and $\pi^{-1}(\bar{H})$ is a subgroup of G containing $\pi^{-1}(\{eN\}) = N$.

These operations are inverse to each other. Given $H \supseteq N$, an element g lies in $\pi^{-1}(\pi(H))$ if and only if $gN = hN$ for some $h \in H$, that is, $g \in hN$ for some $h \in H$; since $hN \subseteq H$, this gives $\pi^{-1}(\pi(H)) = H$. Given $\bar{H} \leqslant G/N$, we have $\pi(\pi^{-1}(\bar{H})) = \bar{H}$ because π is surjective.

(a) If $H_1 \subseteq H_2$, then $\pi(H_1) \subseteq \pi(H_2)$; conversely, if $\pi(H_1) \subseteq \pi(H_2)$, then $H_1 = \pi^{-1}(\pi(H_1)) \subseteq \pi^{-1}(\pi(H_2)) = H_2$. For the index, consider the map from left cosets of H_1 in H_2 to left cosets of H_1/N in H_2/N,

$$hH_1 \mapsto \pi(h)\,(H_1/N) = \pi(hH_1).$$

It is well defined and clearly surjective. It is injective: if $\pi(h)(H_1/N) = \pi(h')(H_1/N)$, then $\pi(h^{-1}h') \in H_1/N$, so $h^{-1}h' \in \pi^{-1}(H_1/N) = H_1$ and $hH_1 = h'H_1$. Hence the two indices are equal.

(b) If $H \trianglelefteq G$, then for $gN \in G/N$ and $hN \in H/N$ we have $(gN)(hN)(gN)^{-1} = ghg^{-1}N \in H/N$, since $ghg^{-1} \in H$. Conversely, if $H/N \trianglelefteq G/N$, then for $g \in G$ and $h \in H$, $\pi(ghg^{-1}) = (gN)(hN)(gN)^{-1} \in H/N$, so $ghg^{-1} \in \pi^{-1}(H/N) = H$. □

Question 3.3.8

Let $G = \mathbb{Z}$ and $N = 12\mathbb{Z}$.

(a) List all subgroups of $G/N = \mathbb{Z}/12$.

(b) For each, find the corresponding subgroup of $\mathbb{Z}$ containing $12\mathbb{Z}$.

(c) Which of these are normal in $\mathbb{Z}$? (Trick question!)

(d) Verify the Third Isomorphism Theorem by checking directly that $(\mathbb{Z}/12\mathbb{Z})/(4\mathbb{Z}/12\mathbb{Z}) \cong \mathbb{Z}/4\mathbb{Z}$.

Solution. **(a)** $\mathbb{Z}/12$ is cyclic of order 12, so its subgroups are the $\langle \bar{d} \rangle$ for the divisors d of 12 (Proposition 2.6.7): $\langle \bar{0} \rangle$, $\langle \bar{6} \rangle$, $\langle \bar{4} \rangle$, $\langle \bar{3} \rangle$, $\langle \bar{2} \rangle$, $\langle \bar{1} \rangle$, of orders 1, 2, 3, 4, 6, 12 respectively.

(b) For $d \mid 12$, an integer g satisfies $\bar{g} \in \langle \bar{d} \rangle$ if and only if $g \equiv kd \pmod{12}$ for some k, which (since $d \mid 12$) holds if and only if $d \mid g$. So $\pi^{-1}(\langle \bar{d} \rangle) = d\mathbb{Z}$, and the correspondence is

$$\langle \bar{0} \rangle \leftrightarrow 12\mathbb{Z}, \qquad \langle \bar{6} \rangle \leftrightarrow 6\mathbb{Z}, \qquad \langle \bar{4} \rangle \leftrightarrow 4\mathbb{Z},$$
$$\langle \bar{3} \rangle \leftrightarrow 3\mathbb{Z}, \qquad \langle \bar{2} \rangle \leftrightarrow 2\mathbb{Z}, \qquad \langle \bar{1} \rangle \leftrightarrow \mathbb{Z}.$$

(c) All of them: every subgroup of the abelian group $\mathbb{Z}$ is normal.
(d) We have $4\mathbb{Z}/12\mathbb{Z} = \{\bar{0}, \bar{4}, \bar{8}\} \cong \mathbb{Z}/3$, so the quotient $(\mathbb{Z}/12\mathbb{Z})/(4\mathbb{Z}/12\mathbb{Z})$ has order $12/3 = 4$. Explicitly, its elements are the cosets $\{\bar{0}, \bar{4}, \bar{8}\}$, $\{\bar{1}, \bar{5}, \bar{9}\}$, $\{\bar{2}, \bar{6}, \bar{10}\}$, $\{\bar{3}, \bar{7}, \bar{11}\}$. This group is cyclic of order 4, generated by the coset of $\bar{1}$, hence isomorphic to $\mathbb{Z}/4\mathbb{Z}$, as the Third Isomorphism Theorem predicts.

3.3.4 The Commutator Subgroup and Abelianization

Two constructions measure how far a group is from being abelian. The center $Z(G)$ (Definition 3.1.13) collects the elements that commute with everything; it is all of G exactly when G is abelian. The second construction goes the other way: it is the smallest normal subgroup whose quotient is abelian.

Definition 3.3.9 (Commutator subgroup and abelianization)

The ***commutator*** of $a, b \in G$ is $[a, b] = aba^{-1}b^{-1}$; it equals e if and only if $ab = ba$. The ***commutator subgroup*** $[G, G]$ is the subgroup generated by all commutators. Since $[a, b]^{-1} = bab^{-1}a^{-1} = [b, a]$, it consists of all finite products

$$[a_1, b_1][a_2, b_2] \cdots [a_k, b_k], \qquad k \geqslant 0,\ a_i, b_i \in G.$$

The quotient $G^{ab} = G/[G, G]$ is the ***abelianization*** of G.

Proposition 3.3.10 (Properties of the commutator subgroup)

Let G be a group.

(a) $[G, G]$ is a normal subgroup of G.
(b) For $N \trianglelefteq G$, the quotient G/N is abelian if and only if $[G, G] \subseteq N$. In particular G^{ab} is abelian, and G is abelian if and only if $[G, G] = \{e\}$.
(c) Every homomorphism $\varphi : G \to A$ to an abelian group A factors uniquely as $\varphi = \bar{\varphi} \circ \pi$, where $\pi : G \to G^{ab}$ is the canonical projection and $\bar{\varphi} : G^{ab} \to A$ is a homomorphism.

Proof. **(a)** For $g \in G$,

$$g[a,b]g^{-1} = (gag^{-1})(gbg^{-1})(gag^{-1})^{-1}(gbg^{-1})^{-1} = [gag^{-1}, gbg^{-1}],$$

as one sees by expanding the right side. Hence the conjugate $g\big(\prod_i [a_i,b_i]\big)g^{-1} = \prod_i g[a_i,b_i]g^{-1}$ of a product of commutators is again a product of commutators, so $g[G,G]g^{-1} \subseteq [G,G]$.

(b) In G/N, $[aN,bN] = [a,b]N$. So G/N is abelian if and only if $[a,b] \in N$ for all $a,b \in G$, which holds if and only if N contains the subgroup generated by the commutators.

(c) Since A is abelian, $\varphi([a,b]) = [\varphi(a),\varphi(b)] = e$, so every commutator lies in $\ker\varphi$, and therefore $[G,G] \subseteq \ker\varphi$. Now apply Theorem 3.3.2. □

So G^{ab} is the largest abelian quotient of G: if G/N is abelian, then $[G,G] \subseteq N$, and by the Third Isomorphism Theorem $G/N \cong G^{ab}/(N/[G,G])$ is a quotient of G^{ab}. For an abelian group, $G^{ab} = G$. We compute the commutator subgroup of S_n in Example 3.4.18, and the abelianization of a free group in Proposition 3.6.17.

3.4 The Symmetric Group

The symmetric group S_n is arguably the most important finite group. By Cayley's theorem, every finite group is isomorphic to a subgroup of some S_n. But S_n is also interesting in its own right: its structure illustrates conjugacy classes, normal subgroups, and simplicity. Symmetric groups connect to linear algebra through the sign homomorphism and determinants, and to Galois theory through their role in permuting the roots of polynomials.

Recall from Chapter 2 that $S_n = \mathrm{Sym}(\{1,\dots,n\})$ has order $n!$, and the cycle notation: $(a_1\ a_2\ \cdots\ a_k)$, for distinct $a_1,\dots,a_k$, is the k-cycle sending $a_1 \mapsto a_2 \mapsto \cdots \mapsto a_k \mapsto a_1$ and fixing everything else; a 2-cycle is a transposition; and two cycles are disjoint if they move no common element. As always, permutations compose right to left: $\sigma\tau$ means τ first, then σ.

3.4.1 Cycle Decomposition

Disjoint cycles c and d commute. Indeed, if c moves x, then c also moves $c(x)$ (otherwise $c(c(x)) = c(x)$ would force $c(x) = x$), so d fixes both x and $c(x)$, and $cd(x) = c(x) = dc(x)$. The same holds with the roles of c and d exchanged, and if neither moves x, then $cd(x) = x = dc(x)$.

Theorem 3.4.1 (Cycle Decomposition Theorem)

Every permutation $\sigma \in S_n$ is a product of disjoint cycles of length at least 2, uniquely up to the order of the factors. These cycles are the restrictions of σ to the orbits of $\langle\sigma\rangle$ on $\{1,\dots,n\}$ that have more than one element.

Proof. **Existence.** Consider the action of $\langle\sigma\rangle$ on $\{1,\dots,n\}$. Fix i. The elements $i, \sigma(i), \sigma^2(i), \dots$ of the finite set $\{1,\dots,n\}$ cannot all be distinct, and the first repetition must be a return to i: if $\sigma^a(i) = \sigma^b(i)$ with $0 \leqslant a < b$, then $\sigma^{b-a}(i) = i$ because σ is injective. Let $k \geqslant 1$ be least with $\sigma^k(i) = i$. Then $i, \sigma(i), \dots, \sigma^{k-1}(i)$ are distinct, and writing any $m \in \mathbb{Z}$ as $m = qk + r$ with $0 \leqslant r < k$ shows $\sigma^m(i) = \sigma^r(i)$. So the orbit of i is $\{i, \sigma(i), \dots, \sigma^{k-1}(i)\}$, and on it σ acts as the k-cycle $(i\ \sigma(i)\ \cdots\ \sigma^{k-1}(i))$.

The orbits partition $\{1,\dots,n\}$ (Proposition 3.2.5). Let $c_1, \dots, c_r$ be the cycles obtained in this way from the orbits with at least two elements; they are disjoint. We check that $\sigma = c_1 \cdots c_r$. If x lies in the orbit of c_j, then $c_j(x) = \sigma(x)$, and both x and $\sigma(x)$ are fixed by the other cycles, so $c_1 \cdots c_r(x) = \sigma(x)$. If $\sigma(x) = x$, every c_j fixes x.

Uniqueness. Suppose $\sigma = d_1 \cdots d_s$ with the d_j disjoint cycles of length at least 2. If $\sigma(x) = x$, no d_j moves x: otherwise $\sigma(x) = d_j(x) \neq x$, since the other cycles fix both x and $d_j(x)$. If $\sigma(x) \neq x$, then some d_j moves x, and the other cycles fix every point that d_j moves; hence $\sigma^m(x) = d_j^m(x)$ for all m. So the set of points moved by d_j is the $\langle\sigma\rangle$-orbit of x, and d_j agrees with σ there; that is, d_j is the cycle of that orbit constructed above. Hence $\{d_1, \dots, d_s\} = \{c_1, \dots, c_r\}$. □

In practice: write down the successive images of 1 under σ until we return to 1; then do the same starting from the smallest number not yet used, and so on.

Example 3.4.2

For the permutation

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 5 & 6 & 4 & 2 & 1 \end{pmatrix}$$

(with $\sigma(i)$ written below i), the successive images of 1 are $1 \mapsto 3 \mapsto 6 \mapsto 1$, and then $2 \mapsto 5 \mapsto 2$, while 4 is fixed. So $\sigma = (1\,3\,6)(2\,5)$.

Definition 3.4.3 (Cycle type)

The ***cycle type*** of $\sigma \in S_n$ is the list of lengths of the cycles in its disjoint cycle decomposition, including a 1-cycle for each fixed point; it is a partition

of n. We write it as $(1^{m_1}2^{m_2}\cdots n^{m_n})$, where m_k is the number of k-cycles, omitting the terms with $m_k = 0$.

Example 3.4.4

In S_6, the permutation $\sigma = (1\,3\,5)(2\,4)$ has cycle type $(1^1 2^1 3^1)$: one fixed point (6), one 2-cycle, and one 3-cycle. The identity has cycle type (1^6), and a 6-cycle has type (6^1).

3.4.2 Conjugacy in S_n

The conjugacy classes of a group (Definition 3.2.12) are fundamental to its structure and to its representation theory. In S_n they have an elegant description.

Lemma 3.4.5 (Conjugating a cycle)

For $\tau \in S_n$ and a cycle $(a_1\, a_2 \cdots a_k)$,

$$\tau\,(a_1\, a_2 \cdots a_k)\,\tau^{-1} = (\tau(a_1)\ \tau(a_2)\ \cdots\ \tau(a_k)).$$

Proof. Write $c = (a_1 \cdots a_k)$. For $i = \tau(a_j)$, we get $\tau c \tau^{-1}(i) = \tau(c(a_j)) = \tau(a_{j+1})$ (indices mod k), the next entry of the cycle on the right. If $i \notin \{\tau(a_1), \ldots, \tau(a_k)\}$, then $\tau^{-1}(i) \notin \{a_1, \ldots, a_k\}$, so c fixes $\tau^{-1}(i)$ and $\tau c \tau^{-1}(i) = i$. □

Theorem 3.4.6 (Conjugacy classes of S_n)

Two permutations in S_n are conjugate if and only if they have the same cycle type.

Proof. **Conjugate ⇒ same type.** Let $\rho = \tau\sigma\tau^{-1}$ and write $\sigma = c_1 \cdots c_r$ as a product of disjoint cycles. Then $\rho = (\tau c_1 \tau^{-1}) \cdots (\tau c_r \tau^{-1})$, and by Lemma 3.4.5 each $\tau c_j \tau^{-1}$ is a cycle of the same length as c_j. These cycles are disjoint because τ is a bijection, so this is the cycle decomposition of ρ, and ρ has the same cycle type as σ.

Same type ⇒ conjugate. Write σ and ρ as products of disjoint cycles, including 1-cycles for fixed points, listing the cycles in order of decreasing length:

$$\sigma = (a_1 \cdots a_{k_1})(b_1 \cdots b_{k_2}) \cdots, \qquad \rho = (c_1 \cdots c_{k_1})(d_1 \cdots d_{k_2}) \cdots.$$

Each of $1, \ldots, n$ appears exactly once in each line, so $\tau(a_i) = c_i$, $\tau(b_j) = d_j$, and so on, defines a permutation τ. By Lemma 3.4.5, $\tau\sigma\tau^{-1} = \rho$. □

Corollary 3.4.7 (Number of conjugacy classes of S_n)

The number of conjugacy classes of S_n equals $p(n)$, the number of partitions of n (ways of writing n as a sum of positive integers, up to order). For example, $p(1) = 1$, $p(2) = 2$, $p(3) = 3$, $p(4) = 5$, $p(5) = 7$, $p(6) = 11$.

Proof. By the theorem, conjugacy classes correspond to cycle types, and every partition $n = k_1 + \cdots + k_\ell$ is the cycle type of some permutation, for instance $(1 \cdots k_1)(k_1{+}1 \cdots k_1{+}k_2)\cdots$. □

Example 3.4.8 (Conjugacy classes of S_3 and S_4)

In S_3 the classes are $\{e\}$, the three transpositions, and the two 3-cycles, of sizes $1 + 3 + 2 = 6$. In S_4:

cycle type	description	size of class
(1^4)	identity	1
$(1^2 2^1)$	transpositions $(i\,j)$	6
(2^2)	$(i\,j)(k\,l)$, two disjoint transpositions	3
$(1^1 3^1)$	3-cycles	8
(4^1)	4-cycles	6

and $1 + 6 + 3 + 8 + 6 = 24$.

NOTE That conjugacy classes of S_n correspond to partitions of n is one reason the representation theory of S_n is so rich: the irreducible representations of S_n are also indexed by partitions of n, and the character table can be computed combinatorially. This is a whole subject in itself; Chapter 14 works out the case $n = 5$.

3.4.3 Generators of S_n

Proposition 3.4.9 (Generators of S_n)

(a) Every k-cycle is a product of $k - 1$ transpositions:

$$(a_1\ a_2\ \cdots\ a_k) = (a_1\ a_k)(a_1\ a_{k-1})\cdots(a_1\ a_2).$$

(b) S_n is generated by the transpositions.

(c) For $n \geqslant 2$, S_n is generated by the $n - 1$ adjacent transpositions $s_i = (i\ \ i{+}1)$, $1 \leqslant i \leqslant n - 1$.

Proof. **(a)** The rightmost factor acts first. Apply the right side to a_1: $(a_1\ a_2)$ sends $a_1 \mapsto a_2$, and the remaining transpositions fix a_2; so $a_1 \mapsto a_2$. Apply it to a_j with $2 \leqslant j \leqslant k-1$: the transpositions $(a_1\ a_2), \dots, (a_1\ a_{j-1})$ fix a_j; then $(a_1\ a_j)$ sends $a_j \mapsto a_1$; then $(a_1\ a_{j+1})$ sends $a_1 \mapsto a_{j+1}$; and the remaining transpositions fix a_{j+1}. So $a_j \mapsto a_{j+1}$. Apply it to a_k: every factor except the last, $(a_1\ a_k)$, fixes a_k, and that one sends $a_k \mapsto a_1$. Elements outside $\{a_1, \dots, a_k\}$ are fixed by both sides.

(b) By the Cycle Decomposition Theorem every permutation is a product of cycles, and by (a) every cycle is a product of transpositions. (The identity is the empty product.)

(c) By (b) it suffices to write each transposition $(i\,j)$, $i < j$, as a product of adjacent transpositions. We induct on $j - i$. If $j = i+1$, then $(i\,j) = s_i$. If $j > i+1$, Lemma 3.4.5 gives

$$(i\ \ i{+}1)\,(i{+}1\ \ j)\,(i\ \ i{+}1) = (i\,j),$$

since $(i\ \ i{+}1)$ sends $i{+}1 \mapsto i$ and fixes j, and $(i\ \ i{+}1)^{-1} = (i\ \ i{+}1)$. By induction $(i{+}1\ \ j)$ is a product of adjacent transpositions, hence so is $(i\,j)$. □

NOTE Part (c) is the mathematics behind the *bubble sort* algorithm. Multiplying σ on the right by s_i swaps the entries in positions i and $i+1$ of the list $\sigma(1), \dots, \sigma(n)$, since $(\sigma s_i)(i) = \sigma(i+1)$. Bubble sort puts the list in order by such swaps, $\sigma s_{i_1} \cdots s_{i_m} = e$, which writes $\sigma = s_{i_m} \cdots s_{i_1}$. Chapter 12 shows that the least number of adjacent transpositions needed is the number of inversions of σ, defined next.

3.4.4 The Sign Homomorphism and the Alternating Group

Definition 3.4.10 (Inversions and sign)

An ***inversion*** of $\sigma \in S_n$ is a pair (i, j) with $1 \leqslant i < j \leqslant n$ and $\sigma(i) > \sigma(j)$. Let $\mathrm{inv}(\sigma)$ be the number of inversions. The ***sign*** of σ is

$$\mathrm{sgn}(\sigma) = (-1)^{\mathrm{inv}(\sigma)} \in \{1, -1\}.$$

A permutation is ***even*** if $\mathrm{sgn}(\sigma) = 1$ and ***odd*** if $\mathrm{sgn}(\sigma) = -1$.

Example 3.4.11

In S_3, the 3-cycle $(1\,2\,3)$ sends $1, 2, 3$ to $2, 3, 1$. Its inversions are $(1, 3)$ and $(2, 3)$, so it is even. For a transposition $(a\,b)$ with $a < b$, the inversions

are (a, b) itself and the pairs (a, m) and (m, b) with $a < m < b$; there are $2(b - a - 1) + 1$ of them, an odd number, so every transposition is odd.

Lemma 3.4.12 (A product formula for the sign)

For $\sigma \in S_n$,

$$\operatorname{sgn}(\sigma) = \prod_{1 \leqslant i < j \leqslant n} \frac{\sigma(j) - \sigma(i)}{j - i}.$$

Proof. As $\{i, j\}$ runs over the two-element subsets of $\{1, \dots, n\}$, so does $\{\sigma(i), \sigma(j)\}$, because σ is a bijection. Hence $\prod_{i<j} |\sigma(j) - \sigma(i)| = \prod_{i<j} (j - i)$, and the product in the lemma has absolute value 1. A factor is negative exactly when (i, j) is an inversion, so the sign of the product is $(-1)^{\operatorname{inv}(\sigma)}$. □

Theorem 3.4.13 (The sign is a homomorphism)

The map $\operatorname{sgn} : S_n \to \{1, -1\}$ is a homomorphism. For $n \geqslant 2$ it is surjective.

Proof. Let $\sigma, \tau \in S_n$. By Lemma 3.4.12,

$$\operatorname{sgn}(\sigma\tau) = \prod_{i<j} \frac{\sigma(\tau(j)) - \sigma(\tau(i))}{j - i} = \prod_{i<j} \frac{\sigma(\tau(j)) - \sigma(\tau(i))}{\tau(j) - \tau(i)} \cdot \prod_{i<j} \frac{\tau(j) - \tau(i)}{j - i}.$$

The second product is $\operatorname{sgn}(\tau)$. In the first, the factor $\frac{\sigma(l) - \sigma(k)}{l - k}$ is unchanged when k and l are exchanged, so it depends only on the set $\{k, l\}$. As $\{i, j\}$ runs over all two-element subsets, so does $\{\tau(i), \tau(j)\}$; hence the first product equals $\prod_{k<l} \frac{\sigma(l) - \sigma(k)}{l - k} = \operatorname{sgn}(\sigma)$. So $\operatorname{sgn}(\sigma\tau) = \operatorname{sgn}(\sigma) \operatorname{sgn}(\tau)$. For $n \geqslant 2$, the transposition $(1\,2)$ has sign -1, so sgn is surjective. □

Corollary 3.4.14 (Parity and transpositions)

(a) If $\sigma = \tau_1 \cdots \tau_r$ is a product of transpositions, then $\operatorname{sgn}(\sigma) = (-1)^r$. Hence the number of transpositions in any such expression of σ is even if σ is even and odd if σ is odd.

(b) A k-cycle has sign $(-1)^{k-1}$.

(c) If $c(\sigma)$ is the number of cycles of σ, counting 1-cycles, then $\operatorname{sgn}(\sigma) = (-1)^{n - c(\sigma)}$.

Proof. (a) Each transposition has sign -1, and sgn is a homomorphism. (b) A k-cycle is a product of $k - 1$ transpositions (Proposition 3.4.9). (c) If the cycles of σ

have lengths $k_1, \dots, k_c$, with $k_1 + \cdots + k_c = n$, then by (b) $\operatorname{sgn}(\sigma) = \prod_i (-1)^{k_i - 1} = (-1)^{n-c}$. □

So the parity of the number of transpositions needed to write σ is well defined, even though the number itself is not: $(1\,2\,3) = (1\,3)(1\,2) = (1\,3)(1\,2)(1\,2)(1\,2)$. This is the nontrivial point; everything else about the sign is bookkeeping.

Definition 3.4.15 (Alternating group)

The ***alternating group*** is $A_n = \ker(\operatorname{sgn}) = \{\sigma \in S_n \mid \operatorname{sgn}(\sigma) = 1\}$, the group of even permutations.

Proposition 3.4.16 (Properties of A_n)

Let $n \geqslant 2$.

(a) $A_n \trianglelefteq S_n$, $[S_n : A_n] = 2$, $|A_n| = n!/2$, and $S_n/A_n \cong \{1, -1\} \cong \mathbb{Z}/2$.

(b) For $n \geqslant 3$, A_n is generated by the 3-cycles. In fact $A_n = \langle (1\,2\,3), (1\,2\,4), \dots, (1\,2\,n) \rangle$.

Proof. **(a)** A_n is the kernel of the surjective homomorphism sgn, so it is normal, and the First Isomorphism Theorem gives $S_n/A_n \cong \{1, -1\}$. Hence $[S_n : A_n] = 2$ and $|A_n| = n!/2$ by Lagrange's theorem.

(b) A 3-cycle is even, since $(a\,b\,c) = (a\,c)(a\,b)$. Conversely, let $\sigma \in A_n$. By Corollary 3.4.14, any expression of σ as a product of transpositions has an even number of factors, so σ is a product of pairs $(a\,b)(c\,d)$. Each pair is a product of 3-cycles: it is e if $\{a, b\} = \{c, d\}$; if the pair shares exactly one element, we may write it as $(a\,b)(b\,d) = (a\,b\,d)$; and if $\{a, b\} \cap \{c, d\} = \emptyset$, then

$$(a\,b)(c\,d) = (a\,b)(b\,c)(b\,c)(c\,d) = (a\,b\,c)(b\,c\,d).$$

So A_n is generated by the 3-cycles.

Now let $H = \langle (1\,2\,k) \mid 3 \leqslant k \leqslant n \rangle$. It suffices to show that every 3-cycle lies in H. Let a, b, c be distinct elements of $\{3, \dots, n\}$. Each identity below can be checked directly, or with Lemma 3.4.5:

- 3-cycles moving both 1 and 2 are $(1\,2\,a)$ and $(1\,a\,2) = (1\,2\,a)^{-1}$;
- 3-cycles moving 1 but not 2 are of the form $(1\,a\,b) = (1\,2\,b)(1\,2\,a)^{-1}$;
- 3-cycles moving 2 but not 1 are of the form $(2\,a\,b) = (1\,2\,b)^{-1}(1\,2\,a)$;
- 3-cycles moving neither are of the form $(a\,b\,c) = (1\,2\,a)(2\,b\,c)(1\,2\,a)^{-1}$, and $(2\,b\,c) \in H$ by the previous case.

The last identity is Lemma 3.4.5, since $(1\,2\,a)$ sends $2 \mapsto a$ and fixes b and c. □

For $n \geqslant 5$, A_n is *simple*; the proof is in Chapter 11 (Theorem 11.4.9). For $n = 4$ this fails: A_4 has a normal subgroup of order 4 (Question 3.4.19 below). The group $A_3 \cong \mathbb{Z}/3$ is simple, but abelian.

NOTE The simplicity of A_n for $n \geqslant 5$ is a historically important result, tied to the non-existence of a formula for the roots of a general quintic. The quadratic formula contains a $\pm\sqrt{\ }$, and the ambiguity of the sign is a $\mathbb{Z}/2 \cong S_2$ permuting the two roots; Cardano's formula for cubics has nested square and cube roots, whose ambiguities combine to an S_3 permuting the three roots. A formula for the roots of a degree 5 equation would need a built-in S_5 symmetry, but the symmetries of any expression built from k-th roots are assembled from cyclic groups $\mathbb{Z}/k$, in the sense that they admit a chain of subgroups, each normal in the next, with abelian quotients. Since A_5 is simple and non-abelian, S_5 admits no such chain. Galois theory makes this precise.

NOTE The sign is intimately related to the determinant (Section 5.4): for an $n \times n$ matrix $A = (a_{ij})$, $\det(A) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)}$, and $\operatorname{sgn}(\sigma)$ is the determinant of the permutation matrix of σ. The sign is what makes the determinant alternating: swapping two rows, or two columns, negates it.

Example 3.4.17 (The converse of Lagrange's theorem fails)

The group A_4 has order $4!/2 = 12$ but no subgroup of order 6. Suppose $H \leqslant A_4$ had order 6, so $[A_4 : H] = 2$, and let g be a 3-cycle. If $g \notin H$, the two cosets of H are H and gH, and g^2H is one of them. It is not gH, since $g^2H = gH$ would give $g \in H$; so $g^2H = H$ and $g^2 \in H$. But then $g = g^4 = (g^2)^2 \in H$, because $g^3 = e$, a contradiction. Hence H contains all eight 3-cycles of A_4 (choose the fixed point in 4 ways, then one of two cyclic orders of the other three), which is impossible since $|H| = 6$.

Example 3.4.18 (The commutator subgroup of S_n)

For $n \geqslant 2$, $[S_n, S_n] = A_n$, and hence $S_n^{ab} \cong \mathbb{Z}/2$. Since $\{1, -1\}$ is abelian, $\operatorname{sgn}([\sigma, \tau]) = 1$, so $[S_n, S_n] \subseteq A_n$. For $n = 2$, $A_2 = \{e\}$ and there is nothing more to show. For $n \geqslant 3$, every 3-cycle is a commutator,

$$[(a\,b), (a\,c)] = (a\,b)(a\,c)(a\,b)^{-1}(a\,c)^{-1} = (a\,b\,c),$$

and the 3-cycles generate A_n (Proposition 3.4.16). So $A_n \subseteq [S_n, S_n]$.

Question 3.4.19

(a) Compute the cycle type and sign of $\sigma = (123)(45)(14)(253)$ in S_5. (First simplify!)

(b) How many elements of S_7 have cycle type $(1^1 2^1 4^1)$?

(c) Let $V = \{e, (12)(34), (13)(24), (14)(23)\} \subseteq S_4$. Prove that V is a subgroup isomorphic to $\mathbb{Z}/2 \times \mathbb{Z}/2$, that $V \trianglelefteq S_4$, and that $S_4/V \cong S_3$.

(d) Show that the normal subgroups of S_4 are $\{e\}$, V, A_4, and S_4.

Solution. **(a)** Multiply the cycles right to left: first (253), then (14), then (45), then (123). Trace each element:

$$1 \xrightarrow{(253)} 1 \xrightarrow{(14)} 4 \xrightarrow{(45)} 5 \xrightarrow{(123)} 5, \quad 2 \xrightarrow{(253)} 5 \xrightarrow{(14)} 5 \xrightarrow{(45)} 4 \xrightarrow{(123)} 4,$$

$$3 \xrightarrow{(253)} 2 \xrightarrow{(14)} 2 \xrightarrow{(45)} 2 \xrightarrow{(123)} 3, \quad 4 \xrightarrow{(253)} 4 \xrightarrow{(14)} 1 \xrightarrow{(45)} 1 \xrightarrow{(123)} 2,$$

$$5 \xrightarrow{(253)} 3 \xrightarrow{(14)} 3 \xrightarrow{(45)} 3 \xrightarrow{(123)} 1.$$

Thus $\sigma = (15)(24)$, of cycle type $(1^1 2^2)$, and $\operatorname{sgn}(\sigma) = (-1)^{5-3} = 1$: σ is even. (Alternatively, the signs of the four factors are $1, -1, -1, 1$, whose product is 1.)

(b) Choose the fixed point (7 ways), then the four elements of the 4-cycle from the remaining six ($\binom{6}{4} = 15$ ways), then arrange them in a 4-cycle ($(4-1)! = 6$ ways, since a k-cycle on a given set of k elements can be written in exactly one way with a chosen element first, followed by an ordering of the other $k-1$). The remaining two elements form the 2-cycle. Total: $7 \cdot 15 \cdot 6 = 630$.

(c) By direct computation, the product of any two distinct non-identity elements of V is the third, for example $(12)(34) \cdot (13)(24) = (14)(23)$, and each non-identity element has order 2. So V is a subgroup in which every element satisfies $x^2 = e$, and $(12)(34) \mapsto (1,0)$, $(13)(24) \mapsto (0,1)$, $(14)(23) \mapsto (1,1)$, $e \mapsto (0,0)$ is an isomorphism $V \to \mathbb{Z}/2 \times \mathbb{Z}/2$, as the two multiplication tables agree. The non-identity elements of V form the conjugacy class of cycle type (2^2) (Theorem 3.4.6), so V is a union of conjugacy classes and hence normal (Proposition 3.2.13).

For the quotient, let S_4 act on the set $P = \{P_1, P_2, P_3\}$ of the three ways to split $\{1,2,3,4\}$ into two pairs,

$$P_1 = \{\{1,2\},\{3,4\}\}, \quad P_2 = \{\{1,3\},\{2,4\}\}, \quad P_3 = \{\{1,4\},\{2,3\}\},$$

by applying σ to every entry. This gives a homomorphism $\rho : S_4 \to \operatorname{Sym}(P) \cong S_3$. The transposition (12) fixes P_1 and swaps P_2 and P_3, and (123) sends

$P_1 \mapsto P_3 \mapsto P_2 \mapsto P_1$. So $\operatorname{Im}\rho$ contains elements of orders 2 and 3, its order is divisible by 6, and ρ is surjective. Each element of V fixes all three P_i; for instance $(1\,2)(3\,4)$ sends $\{1,3\} \mapsto \{2,4\}$ and $\{2,4\} \mapsto \{1,3\}$, so it fixes P_2. Thus $V \subseteq \ker\rho$. By the First Isomorphism Theorem $|\ker\rho| = 24/6 = 4 = |V|$, so $\ker\rho = V$ and $S_4/V \cong S_3$.
(d) A normal subgroup is a union of conjugacy classes containing $\{e\}$, and its order divides 24. The classes of S_4 have sizes $1, 6, 3, 8, 6$ (Example 3.4.8). The sums of 1 with some of $6, 3, 8, 6$ that divide 24 are 1, $1+3=4$, $1+3+8=12$, and 24, each obtained in only one way. The union of size 4 is V, which is a normal subgroup by (c). The union of size 12 consists of the identity, the products of two disjoint transpositions, and the 3-cycles; these are exactly the even permutations (Corollary 3.4.14), so this union is A_4. Hence the normal subgroups of S_4 are $\{e\}$, V, A_4, and S_4. Since $V \subseteq A_4$, V is also a normal subgroup of A_4.

3.5 Exact Sequences

The isomorphism theorems reveal a pattern: kernels and images of homomorphisms are intimately related. Exact sequences are a compact language for these relationships, and they are the starting point of *homological algebra*, a subject that runs through modern mathematics from algebraic topology to representation theory to algebraic geometry. We introduce the basic notions here.

3.5.1 Exact Sequences and Their Meaning

Definition 3.5.1 (Exact sequence)

A sequence of groups and homomorphisms

$$\cdots \longrightarrow G_{i-1} \xrightarrow{\varphi_{i-1}} G_i \xrightarrow{\varphi_i} G_{i+1} \longrightarrow \cdots$$

is ***exact at*** G_i if $\operatorname{Im}(\varphi_{i-1}) = \ker(\varphi_i)$. The sequence is ***exact*** if it is exact at every group other than the two ends (if there are ends).

The condition $\operatorname{Im}(\varphi_{i-1}) = \ker(\varphi_i)$ says: an element of G_i is sent to the identity by φ_i exactly when it comes from G_{i-1}. This is stronger than $\operatorname{Im}(\varphi_{i-1}) \subseteq \ker(\varphi_i)$, which says only that $\varphi_i \circ \varphi_{i-1}$ is the trivial homomorphism.

Example 3.5.2 (Exactness encodes injectivity and surjectivity)

Write $\{e\}$ for the trivial group; the only homomorphisms into or out of it are the trivial ones.

(a) $\{e\} \to A \xrightarrow{\varphi} B$ is exact at A if and only if $\ker(\varphi) = \{e_A\}$, that is, φ is **injective**.

(b) $A \xrightarrow{\varphi} B \to \{e\}$ is exact at B if and only if $\operatorname{Im}(\varphi) = B$, that is, φ is **surjective**.

(c) $\{e\} \to A \xrightarrow{\varphi} B \to \{e\}$ is exact if and only if φ is an **isomorphism**.

3.5.2 Short Exact Sequences

The most important exact sequences are the short ones, which capture quotient groups.

Definition 3.5.3 (Short exact sequence)

A ***short exact sequence*** is an exact sequence of the form

$$\{e\} \longrightarrow N \xrightarrow{\iota} G \xrightarrow{\pi} Q \longrightarrow \{e\}.$$

Exactness means: ι is injective, π is surjective, and $\operatorname{Im}(\iota) = \ker(\pi)$.

The trivial group is usually written 1 for groups in multiplicative notation and 0 for abelian groups in additive notation, so a short exact sequence appears as $1 \to N \to G \to Q \to 1$ or $0 \to A \to B \to C \to 0$.

Proposition 3.5.4 (Short exact sequences and quotients)

If $1 \to N \xrightarrow{\iota} G \xrightarrow{\pi} Q \to 1$ is exact, then $K = \iota(N)$ is a normal subgroup of G isomorphic to N, and π induces an isomorphism $G/K \to Q$, $gK \mapsto \pi(g)$. Conversely, if $K \trianglelefteq G$, then

$$1 \to K \xrightarrow{\text{incl}} G \xrightarrow{\text{proj}} G/K \to 1$$

is exact. So a short exact sequence $1 \to N \to G \to Q \to 1$ exists if and only if G has a normal subgroup isomorphic to N with quotient isomorphic to Q.

Proof. Suppose the sequence is exact. Since ι is injective, it is an isomorphism from N onto $K = \iota(N)$. Exactness at G gives $K = \ker \pi$, which is normal, and since π is surjective the First Isomorphism Theorem gives $G/K \cong Q$ via $gK \mapsto \pi(g)$.

Conversely, for $K \trianglelefteq G$ the inclusion is injective, the projection is surjective, and the image of the inclusion, K, is the kernel of the projection (Theorem 3.1.10). For the last statement, given $K \trianglelefteq G$ and isomorphisms $\alpha : N \to K$ and $\beta : G/K \to Q$, the sequence $1 \to N \xrightarrow{\alpha} G \xrightarrow{\beta \circ \mathrm{proj}} Q \to 1$ is exact. □

When $1 \to N \to G \to Q \to 1$ is exact, we say that G is an ***extension*** of Q by N.

Example 3.5.5 (Fundamental examples of short exact sequences)

(a) **Integers mod n:** $0 \to n\mathbb{Z} \xhookrightarrow{\iota} \mathbb{Z} \xrightarrow{\pi} \mathbb{Z}/n \to 0$, where ι is the inclusion and π is reduction mod n.

(b) **Cyclic groups:** for positive integers m, n, the sequence $0 \to \mathbb{Z}/m \to \mathbb{Z}/mn \to \mathbb{Z}/n \to 0$, where the first map is $k \mapsto nk$ and the second is reduction mod n. The first map is well defined and injective because $nk \equiv nk' \pmod{mn}$ if and only if $k \equiv k' \pmod{m}$; its image is the set of multiples of n in $\mathbb{Z}/mn$, which is the kernel of reduction mod n. For example, $0 \to \mathbb{Z}/2 \to \mathbb{Z}/6 \to \mathbb{Z}/3 \to 0$ and $0 \to \mathbb{Z}/3 \to \mathbb{Z}/6 \to \mathbb{Z}/2 \to 0$.

(c) **Sign:** $1 \to A_n \to S_n \xrightarrow{\mathrm{sgn}} \{\pm 1\} \to 1$ for $n \geqslant 2$. For $n = 3$, with $A_3 \cong \mathbb{Z}/3$ and $\{\pm 1\} \cong \mathbb{Z}/2$, this reads $1 \to \mathbb{Z}/3 \to S_3 \to \mathbb{Z}/2 \to 1$. Together with $0 \to \mathbb{Z}/3 \to \mathbb{Z}/6 \to \mathbb{Z}/2 \to 0$, this shows that the two ends of a short exact sequence do not determine the middle group.

(d) **Determinant:** $1 \to \mathrm{SL}_n(\mathbb{R}) \to \mathrm{GL}_n(\mathbb{R}) \xrightarrow{\det} \mathbb{R}^\times \to 1$ (Example 3.3.4).

(e) **Center:** for any group G, $1 \to Z(G) \to G \to \mathrm{Inn}(G) \to 1$, where the second map is $g \mapsto c_g$ (Example 3.3.4).

(f) **Products:** for any groups N and Q, $1 \to N \xhookrightarrow{\iota} N \times Q \xrightarrow{\pi} Q \to 1$, where $\iota(n) = (n, e)$ and $\pi(n, q) = q$. This is the "trivial" extension.

Example 3.5.6 (A sequence that does not exist)

There is no short exact sequence $1 \to \mathbb{Z}/2 \to S_3 \to \mathbb{Z}/3 \to 1$: it would make the image of $\mathbb{Z}/2$ a normal subgroup of S_3 of order 2, and S_3 has none (Example 3.1.12).

3.5.3 Splitting

When does a short exact sequence "come from" a simpler construction? This leads to the notion of splitting.

Definition 3.5.7 (Split exact sequence)

A short exact sequence $1 \to N \xrightarrow{\iota} G \xrightarrow{\pi} Q \to 1$ ***splits*** if there is a homomorphism $\theta : Q \to G$ with $\pi \circ \theta = \mathrm{id}_Q$. Such a θ is called a ***section*** (or ***splitting***).

Proposition 3.5.8 (Splitting lemma)

If $1 \to N \xrightarrow{\iota} G \xrightarrow{\pi} Q \to 1$ has a section θ, then every element $g \in G$ can be written uniquely as $g = \iota(n)\,\theta(q)$ with $n \in N$ and $q \in Q$.

Proof. **Existence:** Given $g \in G$, let $q = \pi(g)$. Then $\pi(g\,\theta(q)^{-1}) = \pi(g)\,\pi(\theta(q))^{-1} = qq^{-1} = e$, so $g\,\theta(q)^{-1} \in \ker\pi = \iota(N)$. Write $g\,\theta(q)^{-1} = \iota(n)$; then $g = \iota(n)\,\theta(q)$.

Uniqueness: Suppose $\iota(n_1)\,\theta(q_1) = \iota(n_2)\,\theta(q_2)$. Applying π gives $q_1 = q_2$, since $\pi \circ \iota$ is trivial and $\pi \circ \theta = \mathrm{id}_Q$. Cancelling $\theta(q_1)$ gives $\iota(n_1) = \iota(n_2)$, and $n_1 = n_2$ because ι is injective. □

Corollary 3.5.9 (Split sequences of abelian groups)

If $0 \to A \xrightarrow{\iota} B \xrightarrow{\pi} C \to 0$ is a split short exact sequence of abelian groups with section θ, then $A \times C \to B$, $(a, c) \mapsto \iota(a) + \theta(c)$, is an isomorphism. In particular $B \cong A \times C$.

Proof. The map is a homomorphism because B is abelian: $\iota(a + a') + \theta(c + c') = (\iota(a) + \theta(c)) + (\iota(a') + \theta(c'))$. It is bijective by the splitting lemma. □

For non-abelian groups the conclusion fails, because the product $\iota(n)\theta(q)$ need not be multiplicative in the pair (n, q). The correct generalization is the *semidirect product*, studied in Chapter 11: the extensions of Q by N that split are exactly the semidirect products of N and Q. A direct product needs more than a section, namely a retraction onto N.

Proposition 3.5.10 (Retractions and direct products)

For a short exact sequence $1 \to N \xrightarrow{\iota} G \xrightarrow{\pi} Q \to 1$, the following are equivalent:

(a) there is a homomorphism $r : G \to N$ with $r \circ \iota = \mathrm{id}_N$ (a ***retraction***);

(b) there is an isomorphism $\psi : G \to N \times Q$ with $\psi(\iota(n)) = (n, e)$ for all $n \in N$ and $\pi = \mathrm{pr}_Q \circ \psi$, where $\mathrm{pr}_Q(n, q) = q$.

Proof. (b) $\Rightarrow$ (a): Let $r = \mathrm{pr}_N \circ \psi$, where $\mathrm{pr}_N(n, q) = n$. Then $r(\iota(n)) = \mathrm{pr}_N(n, e) = n$.

(a) $\Rightarrow$ (b): Let $\psi(g) = (r(g), \pi(g))$, a homomorphism $G \to N \times Q$ with the two stated properties. It is injective: if $r(g) = e$ and $\pi(g) = e$, then $g = \iota(n)$ for some n by exactness, and $n = r(\iota(n)) = r(g) = e$, so $g = e$. It is surjective: given (n, q), choose g with $\pi(g) = q$, and set $g' = \iota\big(n\, r(g)^{-1}\big)\, g$. Then $\pi(g') = \pi(g) = q$ and $r(g') = n\, r(g)^{-1} r(g) = n$. □

Example 3.5.11 (Split and non-split sequences)

(a) **Splits, but is not a product:** $1 \to A_3 \to S_3 \xrightarrow{\mathrm{sgn}} \{\pm 1\} \to 1$. The homomorphism $\theta : \{\pm 1\} \to S_3$ with $\theta(-1) = (1\,2)$ is a section, since $(1\,2)^2 = e$ and $\mathrm{sgn}(1\,2) = -1$. There is no retraction: by Proposition 3.5.10 it would give $S_3 \cong A_3 \times \{\pm 1\}$, which is abelian, while S_3 is not.

(b) **Does not split:** $0 \to \mathbb{Z}/2 \xrightarrow{\times 2} \mathbb{Z}/4 \to \mathbb{Z}/2 \to 0$, with the second map reduction mod 2. A section $\theta : \mathbb{Z}/2 \to \mathbb{Z}/4$ would satisfy $\theta(1) \equiv 1 \pmod 2$, so $\theta(1) \in \{1, 3\}$, and also $2\theta(1) = \theta(0) = 0$, so $\theta(1) \in \{0, 2\}$. This is impossible. (Alternatively: if it split, Corollary 3.5.9 would give $\mathbb{Z}/4 \cong \mathbb{Z}/2 \times \mathbb{Z}/2$, but $\mathbb{Z}/4$ has an element of order 4 and $\mathbb{Z}/2 \times \mathbb{Z}/2$ does not.)

3.5.4 The Extension Problem

Short exact sequences lead to a fundamental question: given groups N and Q, classify all extensions of Q by N, that is, all groups G fitting into a short exact sequence $1 \to N \to G \to Q \to 1$. This is a deep problem. The split extensions are the semidirect products of Chapter 11; when N is abelian, all extensions, split or not, are classified by cohomology groups $H^2(Q, N)$, which we will not pursue; and since every finite group is built from simple groups by repeated extensions, the problem lies at the heart of the structure theory of finite groups.

Example 3.5.12 (Extensions of $\mathbb{Z}/2$ by $\mathbb{Z}/2$)

Which groups G fit into $1 \to \mathbb{Z}/2 \to G \to \mathbb{Z}/2 \to 1$? Such a G has order 4. There are exactly two groups of order 4 up to isomorphism. Indeed, by Lagrange's theorem the elements of G have order 1, 2, or 4. If some element has order 4, then G is cyclic, $G \cong \mathbb{Z}/4$. Otherwise $x^2 = e$ for all x, so G is abelian ($ab = (ab)^{-1} = b^{-1}a^{-1} = ba$), and for distinct non-identity elements a, b we get $G = \{e, a, b, ab\} \cong \mathbb{Z}/2 \times \mathbb{Z}/2$. Both groups occur:

- $\mathbb{Z}/4$, via $0 \to \mathbb{Z}/2 \xrightarrow{\times 2} \mathbb{Z}/4 \to \mathbb{Z}/2 \to 0$, which does not split;

- $\mathbb{Z}/2 \times \mathbb{Z}/2$, via the product sequence, which splits.

So up to isomorphism of the middle group there are exactly two extensions of $\mathbb{Z}/2$ by $\mathbb{Z}/2$.

Question 3.5.13

(a) Let $n \geqslant 1$. Verify that $0 \to \mathbb{Z} \xrightarrow{\times n} \mathbb{Z} \to \mathbb{Z}/n \to 0$ is exact, where the second map is reduction mod n.

(b) Let $n \geqslant 3$. Show that $1 \to \langle r \rangle \to D_n \to \mathbb{Z}/2 \to 1$ is exact for a suitable surjection $D_n \to \mathbb{Z}/2$, where $\langle r \rangle \cong \mathbb{Z}/n$ is the rotation subgroup. Does it split? Is D_n the direct product $\mathbb{Z}/n \times \mathbb{Z}/2$?

(c) Let G be a group with normal subgroup N. Prove that the sequence $1 \to N \to G \to G/N \to 1$ is exact (with the obvious maps).

(d) Suppose $1 \to A \xrightarrow{f} B \xrightarrow{g} C \to 1$ and $1 \to C \xrightarrow{h} D \xrightarrow{k} E \to 1$ are exact. Prove that $1 \to A \xrightarrow{f} B \xrightarrow{h \circ g} D \xrightarrow{k} E \to 1$ is exact.

Solution. **(a)** Let $\iota(k) = nk$ and let $\pi : \mathbb{Z} \to \mathbb{Z}/n$ be reduction mod n.
Exactness at the first $\mathbb{Z}$*:* $\ker(\iota) = \{k \mid nk = 0\} = \{0\}$, since $n \neq 0$.
Exactness at the second $\mathbb{Z}$*:* $\operatorname{Im}(\iota) = n\mathbb{Z} = \ker(\pi)$.
Exactness at $\mathbb{Z}/n$*:* π is surjective.
(b) Recall (Proposition 2.2.15) that D_n has order $2n$, with rotation subgroup $\langle r \rangle = \{e, r, \ldots, r^{n-1}\}$ of order n and a reflection $s \notin \langle r \rangle$ with $s^2 = e$. The subgroup $\langle r \rangle$ has index 2, so it is normal (Proposition 3.1.11), and $D_n/\langle r \rangle$ is a group of order 2, hence isomorphic to $\mathbb{Z}/2$. Composing the projection with this isomorphism gives a surjection $\pi : D_n \to \mathbb{Z}/2$ with kernel $\langle r \rangle$, so $1 \to \langle r \rangle \to D_n \xrightarrow{\pi} \mathbb{Z}/2 \to 1$ is exact (Proposition 3.5.4).
Does it split? Yes. Define $\theta : \mathbb{Z}/2 \to D_n$ by $\theta(0) = e$ and $\theta(1) = s$; it is a homomorphism because $s^2 = e$. Since $s \notin \langle r \rangle$, $\pi(s) = 1$, so $\pi \circ \theta = \mathrm{id}$. Nevertheless $D_n \ncong \mathbb{Z}/n \times \mathbb{Z}/2$, since D_n is non-abelian ($srs^{-1} = r^{-1} \neq r$ for $n \geqslant 3$); by Proposition 3.5.10 there is no retraction. In the language of Chapter 11, D_n is a semidirect product of $\mathbb{Z}/n$ and $\mathbb{Z}/2$.
(c) Let $\iota : N \to G$ be the inclusion and $\pi : G \to G/N$ the quotient map.
Exactness at N*:* ι is injective.
Exactness at G*:* $\operatorname{Im}(\iota) = N = \ker(\pi)$ by Theorem 3.1.10.
Exactness at G/N*:* π is surjective.
(d) We check exactness at each position.
At A*:* f is injective, by exactness of the first sequence.

At B*:* We need $\operatorname{Im}(f) = \ker(h \circ g)$. By exactness of the first sequence, $\operatorname{Im}(f) = \ker(g)$. If $b \in \ker(g)$, then $(h\circ g)(b) = h(e_C) = e_D$, so $b \in \ker(h\circ g)$. Conversely, if $h(g(b)) = e_D$, then $g(b) \in \ker(h) = \{e_C\}$, since h is injective; so $b \in \ker(g) = \operatorname{Im}(f)$.
At D*:* We need $\operatorname{Im}(h \circ g) = \ker(k)$. Since g is surjective, $\operatorname{Im}(h \circ g) = h(g(B)) = h(C) = \operatorname{Im}(h)$, which equals $\ker(k)$ by exactness of the second sequence.
At E*:* k is surjective, by exactness of the second sequence.

NOTE Exact sequences are the beginning of homological algebra. The ideas extend to longer exact sequences, and to *chain complexes*: sequences in which consecutive composites are trivial but which need not be exact, where the failure of exactness is itself the object of study. In representation theory, exact sequences of representations encode how representations are assembled from smaller ones; in topology, long exact sequences are among the most powerful computational tools.

3.6 Free Groups and Presentations

Every group we have seen so far has been defined concretely: permutations, matrices, integers mod n. There is a fundamentally different approach: describe a group by *generators* and *relations*. This leads to *group presentations*, which are both computationally useful and theoretically important. The key player is the *free group*, the "most general" group on a given set of generators.

3.6.1 Words and Reduced Words

Definition 3.6.1 (Alphabet and words)

Let S be a set (the ***alphabet***). Let $S^{-1} = \{s^{-1} \mid s \in S\}$ be a set of formal symbols in bijection with S and disjoint from it. The elements of $S \cup S^{-1}$ are ***letters***; for a letter x we write x^{-1} for its partner, so that $(s^{-1})^{-1} = s$. A ***word*** in S is a finite sequence of letters,

$$w = x_1 x_2 \cdots x_m, \qquad x_i \in S \cup S^{-1}.$$

The ***empty word***, with $m = 0$, is denoted e. The ***concatenation*** of words $w = x_1 \cdots x_m$ and $w' = y_1 \cdots y_k$ is $ww' = x_1 \cdots x_m y_1 \cdots y_k$; it is associative, with identity e.

Example 3.6.2

If $S = \{a, b\}$, then $ab^{-1}a^{-1}ba$, $aaab$, $b^{-1}b^{-1}a$, and e are words. We abbreviate aaa as a^3 and $a^{-1}a^{-1}$ as a^{-2}.

We want words to form a group, but aa^{-1} and e are different words that "should" be equal. So we cancel adjacent inverse letters.

Definition 3.6.3 (Reduction)

An ***elementary reduction*** of a word deletes two adjacent letters of the form xx^{-1}, for a letter x. A word is ***reduced*** if no elementary reduction applies, that is, no letter is immediately followed by its inverse.

Example 3.6.4

The word $ab^{-1}ba^{-1}a$ can be reduced in two ways:

$$ab^{-1}ba^{-1}a \to aa^{-1}a \to a \qquad \text{and} \qquad ab^{-1}ba^{-1}a \to ab^{-1}b \to a.$$

The first path deletes $b^{-1}b$ and then aa^{-1}; the second deletes $a^{-1}a$ and then $b^{-1}b$. Both end at the reduced word a.

Theorem 3.6.5 (Unique reduced form)

Every word u can be turned into a reduced word by a finite sequence of elementary reductions, and the reduced word obtained does not depend on the sequence chosen.

Proof. **Existence.** Each elementary reduction shortens the word by two letters, so repeated reductions stop after finitely many steps, at a reduced word.

Uniqueness. Let W be the set of reduced words. For each letter x define $\lambda_x : W \to W$ by

$$\lambda_x(w) = \begin{cases} xw & \text{if } w \text{ does not begin with } x^{-1}, \\ w' & \text{if } w = x^{-1}w'. \end{cases}$$

In both cases the result is reduced. We claim $\lambda_x \circ \lambda_{x^{-1}} = \mathrm{id}_W$. If w does not begin with x, then $\lambda_{x^{-1}}(w) = x^{-1}w$, which begins with x^{-1}, so $\lambda_x(x^{-1}w) = w$. If $w = xw'$, then $\lambda_{x^{-1}}(w) = w'$, and w' does not begin with x^{-1} because w is reduced, so $\lambda_x(w') = xw' = w$.

For an arbitrary word $u = x_1 \cdots x_m$, put $\lambda_u = \lambda_{x_1} \circ \cdots \circ \lambda_{x_m}$ (and $\lambda_e = \mathrm{id}_W$). If u' is obtained from u by deleting an adjacent pair xx^{-1}, then $\lambda_{u'} = \lambda_u$, because

$\lambda_x \circ \lambda_{x^{-1}} = \mathrm{id}_W$. If $w = x_1 \cdots x_m$ is reduced, then $\lambda_w(e) = w$: by induction on m, $\lambda_{x_2 \cdots x_m}(e) = x_2 \cdots x_m$, which does not begin with x_1^{-1} because w is reduced, so applying λ_{x_1} gives w.

Now suppose some sequence of elementary reductions takes u to a reduced word w. Then $\lambda_u = \lambda_w$, so $w = \lambda_w(e) = \lambda_u(e)$, which depends only on u. □

Write $\mathrm{red}(u)$ for the reduced word obtained from u. If u can be reduced to u' (by any sequence of elementary reductions), then $\mathrm{red}(u) = \mathrm{red}(u')$, since a reduction sequence from u' to a reduced word extends one from u to u'. In particular, for any words u, v,

$$\mathrm{red}(uv) = \mathrm{red}(\mathrm{red}(u)\, v) = \mathrm{red}(u\ \mathrm{red}(v)).$$

In a reduced word, a maximal block of consecutive letters from $\{s, s^{-1}\}$ consists of copies of s only or of s^{-1} only. Grouping such blocks, every reduced word can be written uniquely as

$$s_1^{m_1} s_2^{m_2} \cdots s_k^{m_k}, \qquad k \geqslant 0,\ s_j \in S,\ s_j \neq s_{j+1},\ m_j \in \mathbb{Z} \setminus \{0\},$$

and conversely every such expression, written out letter by letter, is a reduced word. For $S = \{a_1, \ldots, a_n\}$ this is the form $a_{i_1}^{m_1} \cdots a_{i_k}^{m_k}$ with $i_j \neq i_{j+1}$.

3.6.2 The Free Group

Definition 3.6.6 (Free group)

The ***free group on*** S, denoted $F(S)$, is the set of reduced words in S with the operation

$$w \cdot w' = \mathrm{red}(ww').$$

For $S = \{a_1, \ldots, a_n\}$ we write F_n, the free group on n generators.

Proposition 3.6.7 (The free group is a group)

The set $F(S)$ is a group under this operation. Its identity is the empty word, and the inverse of $x_1 \cdots x_m$ is $x_m^{-1} \cdots x_1^{-1}$.

Proof. The product of two reduced words is reduced by definition.

Associativity: By the remark after Theorem 3.6.5,

$$\begin{aligned}(w_1 \cdot w_2) \cdot w_3 &= \mathrm{red}(\mathrm{red}(w_1 w_2)\, w_3) = \mathrm{red}(w_1 w_2 w_3) \\ &= \mathrm{red}(w_1\ \mathrm{red}(w_2 w_3)) = w_1 \cdot (w_2 \cdot w_3).\end{aligned}$$

Identity: $e \cdot w = \mathrm{red}(w) = w$ since w is already reduced; similarly $w \cdot e = w$.

Inverses: Let $w = x_1 \cdots x_m$ be reduced and $w^{-1} = x_m^{-1} \cdots x_1^{-1}$. This word is reduced: an adjacent pair $x_{i+1}^{-1} x_i^{-1}$ cancels only if $x_i^{-1} = x_{i+1}$, and then $x_i x_{i+1}$ would cancel in w. In ww^{-1} the middle pair $x_m x_m^{-1}$ cancels, then $x_{m-1} x_{m-1}^{-1}$, and so on, until we reach e. So $w \cdot w^{-1} = e$, and likewise $w^{-1} \cdot w = e$. □

If $x_1 \cdots x_m$ is reduced, then it is the product $x_1 \cdot x_2 \cdots x_m$ in $F(S)$ of its one-letter words, and the one-letter word s^{-1} is the inverse of s. So $F(S)$ is generated by S (identifying s with the one-letter word), and the grouped form above expresses each element as a product of powers of generators. Renaming the letters shows that $F(S) \cong F(T)$ whenever there is a bijection $S \to T$.

Example 3.6.8 (Free group on one generator)

$F_1 = F(\{a\})$ consists of the reduced words $\ldots, a^{-2}, a^{-1}, e, a, a^2, \ldots$, that is, $\{a^m \mid m \in \mathbb{Z}\}$, with $a^m \cdot a^k = a^{m+k}$. So $F_1 \cong \mathbb{Z}$.

Example 3.6.9 (Free group on two generators)

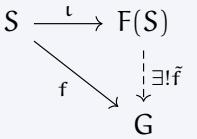
$F_2 = F(\{a, b\})$ is much more interesting. Its elements include $e, a, b, a^{-1}, b^{-1}, ab, ba, a^2, ab^{-1}, aba, bab^{-1}a^{-1}, \ldots$, infinitely many distinct reduced words. Since ab and ba are different reduced words, $ab \neq ba$, and F_2 is non-abelian. In fact F_2 contains subgroups isomorphic to F_n for every n; Question 3.6.22 below proves this.

3.6.3 The Universal Property

The free group is "free" because no relations hold among its generators beyond those forced by the group axioms. This is captured by a *universal property*.

Theorem 3.6.10 (Universal property of free groups)

Let $\iota : S \to F(S)$ send s to the one-letter word s. For any group G and any function $f : S \to G$, there is a unique homomorphism $\tilde{f} : F(S) \to G$ with $\tilde{f} \circ \iota = f$.



Proof. **Existence.** Extend f to letters by $f(s^{-1}) = f(s)^{-1}$, and define Φ on arbitrary words by $\Phi(x_1 \cdots x_m) = f(x_1) \cdots f(x_m)$, with $\Phi(e) = e$. Then $\Phi(uv) = \Phi(u)\Phi(v)$ for all words u, v. An elementary reduction does not change Φ, because

$f(x)f(x^{-1}) = e$ for every letter x; hence $\Phi(u) = \Phi(\mathrm{red}(u))$. Let $\tilde{f}$ be the restriction of Φ to reduced words. Then

$$\tilde{f}(w \cdot w') = \Phi(\mathrm{red}(ww')) = \Phi(ww') = \Phi(w)\Phi(w') = \tilde{f}(w)\tilde{f}(w'),$$

so $\tilde{f}$ is a homomorphism, and $\tilde{f}(s) = f(s)$.

Uniqueness. A homomorphism $\varphi : F(S) \to G$ with $\varphi(s) = f(s)$ for all $s \in S$ satisfies $\varphi(s^{-1}) = f(s)^{-1}$, and since $x_1 \cdots x_m = x_1 \cdot x_2 \cdots x_m$ in $F(S)$,

$$\varphi(x_1 \cdots x_m) = \varphi(x_1) \cdots \varphi(x_m) = f(x_1) \cdots f(x_m) = \tilde{f}(x_1 \cdots x_m). \qquad \square$$

NOTE The universal property says that to define a homomorphism *out of* $F(S)$, it suffices to say where the generators go, and there are no constraints: any choice of images in any group G extends uniquely. This is why $F(S)$ is called "free." Chapter 8 studies universal properties of this kind in general.

Corollary 3.6.11 (Every group is a quotient of a free group)

If elements $g_s \in G$ ($s \in S$) generate G, then the homomorphism $F(S) \to G$ with $s \mapsto g_s$ is surjective, and $G \cong F(S)/K$ for its kernel K.

Proof. The image is a subgroup containing every g_s, so it is all of G. Apply the First Isomorphism Theorem. $\square$

So F_n is the "largest" group with n generators: every group generated by n elements is a quotient of F_n.

3.6.4 Group Presentations

To describe G as a quotient $F(S)/K$, we would like to specify K by a few of its elements. Since K is normal, it then also contains all conjugates of those elements.

Proposition 3.6.12 (Normal closure)

Let R be a subset of a group F. The set $\langle\langle R \rangle\rangle$ of all finite products

$$g_1 r_1^{\epsilon_1} g_1^{-1}\, g_2 r_2^{\epsilon_2} g_2^{-1} \cdots g_k r_k^{\epsilon_k} g_k^{-1}, \qquad k \geqslant 0,\ g_i \in F,\ r_i \in R,\ \epsilon_i = \pm 1,$$

is the smallest normal subgroup of F containing R: it is a normal subgroup containing R, and every normal subgroup containing R contains it. It is called the ***normal closure*** of R.

Proof. The set contains e (take $k = 0$) and R; it is closed under products by construction, and under inverses because $(g r^{\epsilon} g^{-1})^{-1} = g r^{-\epsilon} g^{-1}$ and the inverse

of a product is the product of the inverses in reverse order. It is normal because $h(grg^{-1})h^{-1} = (hg)r(hg)^{-1}$, so a conjugate of such a product is again such a product. Finally, a normal subgroup containing R contains every $gr^{\pm 1}g^{-1}$, hence every such product. □

Definition 3.6.13 (Group presentation)

Let S be a set and $R \subseteq F(S)$ a set of words. The group ***presented*** by generators S and relations R is

$$\langle S \mid R \rangle = F(S)/\langle\langle R \rangle\rangle.$$

A ***presentation*** of a group G is an isomorphism $G \cong \langle S \mid R \rangle$. The group is ***finitely presented*** if it has a presentation with S and R finite; we then write $\langle s_1, \ldots, s_n \mid r_1, \ldots, r_k \rangle$. A relation is often written as an equation: $u = v$ stands for the relation uv^{-1}.

In $\langle S \mid R \rangle$, write $\bar{s}$ for the image of $s \in S$. The $\bar{s}$ generate the group, and every relation $r \in R$ becomes the identity: $\bar{r} = e$. The presented group is the largest group with these properties, in the following sense.

Proposition 3.6.14 (Mapping out of a presentation)

Let G be a group and $f : S \to G$ a function such that $\tilde{f}(r) = e$ for every $r \in R$ (we say the relations *hold* in G). Then there is a unique homomorphism $\bar{f} : \langle S \mid R \rangle \to G$ with $\bar{f}(\bar{s}) = f(s)$ for all $s \in S$. It is surjective if and only if the elements $f(s)$ generate G, and it is an isomorphism if and only if, in addition, $\ker \tilde{f} = \langle\langle R \rangle\rangle$.

Proof. The kernel of $\tilde{f} : F(S) \to G$ is a normal subgroup containing R, so it contains $\langle\langle R \rangle\rangle$ (Proposition 3.6.12). By Theorem 3.3.2, $\tilde{f}$ factors through a homomorphism $\bar{f} : F(S)/\langle\langle R \rangle\rangle \to G$ with the same image as $\tilde{f}$, namely the subgroup generated by the $f(s)$, and with kernel $\ker \tilde{f}/\langle\langle R \rangle\rangle$. This kernel is trivial exactly when $\ker \tilde{f} = \langle\langle R \rangle\rangle$. Uniqueness holds because the $\bar{s}$ generate $\langle S \mid R \rangle$. □

This gives the standard method for proving that a group G has a given presentation. First, choose generators of G and check that the relations hold in G; this gives a surjection $\langle S \mid R \rangle \twoheadrightarrow G$. Second, show that the relations are *complete*. When G is finite, it suffices to show that $\langle S \mid R \rangle$ has at most $|G|$ elements, usually by using the relations to rewrite every element in a normal form; a surjection between finite sets of the same size is a bijection.

Example 3.6.15 (Presentations of familiar groups)

(a) **Cyclic groups:** $\mathbb{Z}/n \cong \langle a \mid a^n \rangle$. The relation holds for the generator $1 \in \mathbb{Z}/n$, giving a surjection. In $\langle a \mid a^n \rangle$ every element is a power $\bar{a}^k$, and $\bar{a}^n = e$, so there are at most n elements $\bar{a}^0, \ldots, \bar{a}^{n-1}$.

(b) **Infinite cyclic group:** $\mathbb{Z} \cong F_1 = \langle a \mid \rangle$, with no relations.

(c) **Klein four-group:** $\mathbb{Z}/2 \times \mathbb{Z}/2 \cong \langle a, b \mid a^2, b^2, aba^{-1}b^{-1} \rangle$. The relations hold for $(1,0)$ and $(0,1)$, which generate. In the presented group $\bar{a}\bar{b} = \bar{b}\bar{a}$ and $\bar{a}^2 = \bar{b}^2 = e$, so every element is $\bar{a}^i\bar{b}^j$ with $i, j \in \{0,1\}$: at most 4 elements.

(d) **The symmetric group S_3:** $S_3 \cong \langle s_1, s_2 \mid s_1^2, s_2^2, (s_1s_2)^3 \rangle$, with $s_1 = (1\,2)$ and $s_2 = (2\,3)$. This is the case $n = 3$ of Theorem 3.6.18 below.

Proposition 3.6.16 (Presentation of the dihedral group)
For $n \geqslant 3$, $D_n \cong \langle r, s \mid r^n, s^2, srs^{-1}r \rangle$. (The last relation says $srs^{-1} = r^{-1}$.)

Proof. By Proposition 2.2.15, D_n has order $2n$ and is generated by a rotation r and a reflection s with $r^n = s^2 = e$ and $srs^{-1} = r^{-1}$. So the relations hold in D_n, and Proposition 3.6.14 gives a surjection from the presented group P onto D_n. In P, the relations give $\bar{s}\bar{r} = \bar{r}^{-1}\bar{s}$, hence $\bar{s}\bar{r}^k = \bar{r}^{-k}\bar{s}$ for all k, and also $\bar{s}^{-1} = \bar{s}$ and $\bar{r}^{-1} = \bar{r}^{n-1}$. So any product of the generators and their inverses can be rewritten by moving each $\bar{s}$ to the right, in the form $\bar{r}^k\bar{s}^\varepsilon$ with $0 \leqslant k < n$ and $\varepsilon \in \{0,1\}$. Thus $|P| \leqslant 2n = |D_n|$, and the surjection is an isomorphism. □

Proposition 3.6.17 (Free abelian groups)
$\mathbb{Z}^n \cong \langle a_1, \ldots, a_n \mid [a_i, a_j] \text{ for } 1 \leqslant i, j \leqslant n \rangle$, where $[a, b] = aba^{-1}b^{-1}$. Moreover $\langle\langle [a_i, a_j] \rangle\rangle = [F_n, F_n]$, so $\mathbb{Z}^n$ is the abelianization of F_n.

Proof. Let P be the presented group and $\epsilon_i \in \mathbb{Z}^n$ the element with 1 in position i and 0 elsewhere. Since $\mathbb{Z}^n$ is abelian, the relations hold for $a_i \mapsto \epsilon_i$, and the ϵ_i generate $\mathbb{Z}^n$, so we get a surjection $\varphi : P \to \mathbb{Z}^n$. In P the generators $\bar{a}_i$ commute pairwise, hence so do their inverses, and therefore any two products of generators and their inverses commute: P is abelian. Collecting powers, every element of P is $\bar{a}_1^{m_1} \cdots \bar{a}_n^{m_n}$ for some $m_i \in \mathbb{Z}$, and φ sends it to $(m_1, \ldots, m_n)$. So φ is injective, hence an isomorphism.

For the second statement, $N = \langle\langle [a_i, a_j] \rangle\rangle$ is contained in $[F_n, F_n]$, which is a normal subgroup containing the $[a_i, a_j]$ (Proposition 3.3.10). Conversely $F_n/N = P$ is abelian, so $[F_n, F_n] \subseteq N$ by Proposition 3.3.10(b). Hence $F_n^{ab} = F_n/[F_n, F_n] = P \cong \mathbb{Z}^n$. □

The symmetric group has a presentation in terms of the adjacent transpositions $s_i = (i\ \ i{+}1)$, which generate it by Proposition 3.4.9.

Theorem 3.6.18 (Presentation of the symmetric group)
For $n \geqslant 1$, let G_n be the group with generators $t_1, \dots, t_{n-1}$ and relations

$$t_i^2 = e, \qquad t_i t_{i+1} t_i = t_{i+1} t_i t_{i+1}, \qquad t_i t_j = t_j t_i \text{ for } |i-j| \geqslant 2.$$

Then $t_i \mapsto s_i = (i\ \ i{+}1)$ defines an isomorphism $G_n \to S_n$. Equivalently,

$$S_n \cong \langle t_1, \dots, t_{n-1} \mid t_i^2,\ (t_i t_{i+1})^3,\ (t_i t_j)^2 \text{ for } |i-j| \geqslant 2 \rangle;$$

these relations are the ***Coxeter relations***.

Proof. Given $t_i^2 = e$, the relation $(t_i t_{i+1})^3 = e$ is equivalent to $t_i t_{i+1} t_i = t_{i+1} t_i t_{i+1}$, and $(t_i t_j)^2 = e$ to $t_i t_j = t_j t_i$; so the two lists of relations define the same group. The relations hold in S_n: $s_i^2 = e$; $s_i s_{i+1} = (i\ \ i{+}1\ \ i{+}2)$ has order 3; and s_i, s_j are disjoint for $|i-j| \geqslant 2$, so they commute. Since the s_i generate S_n, Proposition 3.6.14 gives a surjection $G_n \twoheadrightarrow S_n$. It remains to show $|G_n| \leqslant n!$, by induction on n. For $n = 1$, G_1 has no generators and is trivial.

Let $n \geqslant 2$ and let H be the subgroup of G_n generated by $t_1, \dots, t_{n-2}$. These elements satisfy the relations defining G_{n-1}, so Proposition 3.6.14 gives a surjection $G_{n-1} \twoheadrightarrow H$, and $|H| \leqslant (n-1)!$ by induction. Put $w_n = e$ and $w_k = t_{n-1} t_{n-2} \cdots t_k$ for $1 \leqslant k \leqslant n-1$, and let $U = Hw_1 \cup \cdots \cup Hw_n$, a union of n right cosets of H. We claim that $Ut_i \subseteq U$ for every i. Fix k and i.

- If $i = k-1$, then $w_k t_{k-1} = w_{k-1}$.
- If $i = k$ (so $k \leqslant n-1$), then $w_k t_k = t_{n-1} \cdots t_{k+1} t_k t_k = w_{k+1}$.
- If $i \leqslant k-2$, then t_i commutes with each of $t_{n-1}, \dots, t_k$, so $w_k t_i = t_i w_k$, and $t_i \in H$; hence $Hw_k t_i = Hw_k$.
- If $i \geqslant k+1$, then $w_k = t_{n-1} \cdots t_{i+1}\, t_i t_{i-1}\, t_{i-2} \cdots t_k$. Since t_i commutes with $t_{i-2}, \dots, t_k$, and t_{i-1} commutes with $t_{n-1}, \dots, t_{i+1}$,

$$\begin{aligned} w_k t_i &= t_{n-1} \cdots t_{i+1}\, (t_i t_{i-1} t_i)\, t_{i-2} \cdots t_k \\ &= t_{n-1} \cdots t_{i+1}\, (t_{i-1} t_i t_{i-1})\, t_{i-2} \cdots t_k = t_{i-1} w_k, \end{aligned}$$

  and $t_{i-1} \in H$ since $i - 1 \leqslant n-2$; hence $Hw_k t_i = Hw_k$.

So $Hw_k t_i$ is again one of the cosets Hw_j, proving the claim. Since $t_i^{-1} = t_i$, every element of G_n is a product $t_{i_1} \cdots t_{i_m}$, and starting from $e \in Hw_n \subseteq U$ and multiplying on the right by one generator at a time, we stay in U. Hence $G_n = U$, and $|G_n| \leqslant n\,|H| \leqslant n!$. So the surjection $G_n \twoheadrightarrow S_n$ is an isomorphism. □

Chapter 12 continues the study of presentations, with Cayley graphs, normal forms, the braid group, and $\mathrm{SL}_2(\mathbb{Z})$.

Proposition 3.6.19 (Every group has a presentation)

For every group G there are S and R with $G \cong \langle S \mid R \rangle$.

Proof. Take $S = G$, regarded as a set, and $f = \mathrm{id} : S \to G$. By Corollary 3.6.11, $\tilde{f} : F(S) \to G$ is surjective. Let $R = \ker \tilde{f}$. Since R is already normal, $\langle\langle R \rangle\rangle = R$, and $G \cong F(S)/R = \langle S \mid R \rangle$ by the First Isomorphism Theorem. □

This presentation is enormous, but it exists. Finding small presentations is an art.

3.6.5 The Word Problem

Presentations raise a fundamental computational question.

Definition 3.6.20 (Word problem)

Given a presentation $G = \langle S \mid R \rangle$ and two words w_1, w_2 in S, the ***word problem*** asks whether w_1 and w_2 represent the same element of G; equivalently, whether $w_1 w_2^{-1} \in \langle\langle R \rangle\rangle$.

The following theorem, due to Novikov (1955) and Boone (1959), is beyond the scope of this book; a proof can be found in J. J. Rotman, *An Introduction to the Theory of Groups*, 4th ed., Chapter 12.

Theorem 3.6.21 (Novikov–Boone)

There exists a finitely presented group $G = \langle S \mid R \rangle$ for which the word problem is **undecidable**: no algorithm can determine, for every pair of words, whether they represent the same element of G.

NOTE This is a sobering result. Group presentations look like a convenient computational tool, but they are so expressive that they can encode arbitrary computation: the word problem for a suitable presentation can simulate a Turing machine. This does not make presentations useless. For many specific groups, such as S_n, D_n, and free groups, the word problem is decidable; but there is no algorithm that works for all finitely presented groups.

Question 3.6.22

(a) Prove that F_2 contains a subgroup isomorphic to F_3. More generally, show that F_2 contains a subgroup isomorphic to F_n for every $n \geqslant 1$.
(b) Find a presentation for the quaternion group $Q_8 = \{\pm 1, \pm i, \pm j, \pm k\}$, with $i^2 = j^2 = k^2 = -1$, $ij = k$, $jk = i$, $ki = j$.
(c) Prove that $\langle a, b \mid a^2, b^3, (ab)^2 \rangle \cong S_3$.

Solution. **(a)** Let $F_2 = F(\{a, b\})$, fix $n \geqslant 1$, and let $F_n = F(\{x_0, \dots, x_{n-1}\})$. By the universal property there is a homomorphism $\varphi : F_n \to F_2$ with $\varphi(x_k) = b^k a b^{-k}$. We show that φ is injective; then $F_n \cong \operatorname{Im} \varphi$. For $n = 3$ this is the subgroup $\langle a,\ bab^{-1},\ b^2ab^{-2} \rangle$.
Let $u \neq e$ be a reduced word in the x_k, written in grouped form $u = x_{k_1}^{m_1} x_{k_2}^{m_2} \cdots x_{k_r}^{m_r}$ with $r \geqslant 1$, $k_j \neq k_{j+1}$, and $m_j \neq 0$. Since $\varphi(x_k^m) = (b^k a b^{-k})^m = b^k a^m b^{-k}$,

$$\varphi(u) = b^{k_1} a^{m_1} b^{k_2 - k_1} a^{m_2} b^{k_3 - k_2} \cdots b^{k_r - k_{r-1}} a^{m_r} b^{-k_r}.$$

Every exponent of a is nonzero, and every interior exponent $k_{j+1} - k_j$ of b is nonzero. Deleting the outer factors b^{k_1} and b^{-k_r} if their exponents are zero, we obtain a grouped form in which adjacent letters differ and all exponents are nonzero. This is a reduced word, and it contains a^{m_1}, so it is not empty. Hence $\varphi(u) \neq e$, $\ker \varphi = \{e\}$, and φ is injective.
(b) We show $Q_8 \cong \langle i, j \mid i^4,\ i^2 j^{-2},\ iji^{-1}j \rangle$; the relations say $i^4 = e$, $i^2 = j^2$, and $iji^{-1} = j^{-1}$.
The relations hold in Q_8: $i^4 = (-1)^2 = 1$; $i^2 = -1 = j^2$; and $iji^{-1} = ij(-i) = -ki = -j = j^{-1}$. Also i and j generate Q_8, since $k = ij$, $-1 = i^2$, and $-x = x \cdot i^2$. So we get a surjection from the presented group P onto Q_8.
P *has at most 8 elements:* In P (writing i, j for their images), $j^4 = (j^2)^2 = (i^2)^2 = e$, so $i^{-1} = i^3$ and $j^{-1} = j^3$, and every element is a product of nonnegative powers of i and j. The relation $ij = j^{-1}i$ gives $ij^b = j^{-b}i$, so we can move every i to the right: every element is $j^b i^a$ with $a, b \geqslant 0$. Using $i^2 = j^2$ and $j^4 = e$, we may take $a \in \{0, 1\}$ and $b \in \{0, 1, 2, 3\}$. So $|P| \leqslant 8 = |Q_8|$, and $P \cong Q_8$.
(c) Let $P = \langle a, b \mid a^2, b^3, (ab)^2 \rangle$, and write a, b for the images of the generators in P.
A surjection onto S_3: Send $a \mapsto (1\,2)$ and $b \mapsto (1\,2\,3)$. The relations hold: $(1\,2)^2 = e$, $(1\,2\,3)^3 = e$, and $(1\,2)(1\,2\,3) = (2\,3)$, whose square is e. The subgroup generated by $(1\,2)$ and $(1\,2\,3)$ contains elements of orders 2 and 3,

so its order is divisible by 6, and it is all of S_3. By Proposition 3.6.14 we get a surjection $P \twoheadrightarrow S_3$.
P *has at most 6 elements:* From $abab = e$ and $a^{-1} = a$ we get $aba^{-1} = aba = b^{-1}$, so $ab = b^{-1}a$ and hence $ab^k = b^{-k}a$ for all k. Moving every a to the right, every element of P is $b^k a^\varepsilon$ with $0 \leqslant k \leqslant 2$ and $\varepsilon \in \{0, 1\}$. So $|P| \leqslant 6$, and $P \cong S_3$.

PART II

Linear Algebra

CHAPTER 4

Rings, Fields, and Vector Spaces

This chapter marks our transition from groups to linear algebra. Vector spaces are among the most tractable algebraic structures: finite-dimensional vector spaces over a fixed field are classified up to isomorphism by a single integer, their dimension, whereas nothing so simple is available for groups. Before we can speak of vector spaces we need their scalars, so we begin with rings and fields, their homomorphisms, polynomial rings and the characteristic of a field. We then develop vector spaces, linear independence, bases and dimension.

4.1 Rings and Fields

A group has one operation. The familiar number systems $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ have two, addition and multiplication, linked by the distributive law. Rings and fields axiomatize this situation.

Definition 4.1.1 (Ring)

A ***ring*** is a set R with two binary operations, addition $(a, b) \mapsto a + b$ and multiplication $(a, b) \mapsto ab$, such that:

(a) $(R, +)$ is an abelian group, whose identity element we write 0;

(b) multiplication is associative, $a(bc) = (ab)c$, and has an identity element $1 \in R$: $1a = a1 = a$ for all $a \in R$;

(c) multiplication distributes over addition: $a(b + c) = ab + ac$ and $(a + b)c = ac + bc$ for all $a, b, c \in R$.

The ring is ***commutative*** if $ab = ba$ for all $a, b \in R$.

Example 4.1.2

The integers $\mathbb{Z}$, the rationals $\mathbb{Q}$, the reals $\mathbb{R}$ and the complex numbers $\mathbb{C}$ are commutative rings under the usual operations. Most rings in this book are

commutative. The main exceptions are the rings of $n \times n$ matrices and of linear operators on a vector space, which we meet in Chapter 5.

Proposition 4.1.3 (First consequences of the axioms)

Let R be a ring and $a, b \in R$. Then:

(a) $0a = a0 = 0$;

(b) $(-a)b = a(-b) = -(ab)$; in particular $(-1)a = -a$;

(c) $(-a)(-b) = ab$.

Proof. (a) By distributivity, $a0 = a(0+0) = a0 + a0$. Adding $-(a0)$ to both sides gives $0 = a0$. The same argument gives $0a = 0$.

(b) We have $ab + (-a)b = (a + (-a))b = 0b = 0$, so $(-a)b$ is the additive inverse of ab. Likewise $ab + a(-b) = a(b + (-b)) = 0$. Taking $a = 1$ gives $(-1)b = -b$.

(c) By (b) twice, $(-a)(-b) = -\big(a(-b)\big) = -\big(-(ab)\big) = ab$. □

Proposition 4.1.4 (The zero ring)

In a ring R we have $1 = 0$ if and only if $R = \{0\}$.

Proof. If $1 = 0$, then every $a \in R$ satisfies $a = a1 = a0 = 0$. Conversely, if $R = \{0\}$ then $1 = 0$. □

The set $\{0\}$ with $0 + 0 = 0 \cdot 0 = 0$ is a ring, the ***zero ring***. By the proposition it is the only ring in which $1 = 0$.

Definition 4.1.5 (Units)

An element a of a ring R is a ***unit*** if there is $b \in R$ with $ab = ba = 1$. Such a b is unique: if $ab' = b'a = 1$ as well, then $b = b(ab') = (ba)b' = b'$. We call it the ***inverse*** of a and write it a^{-1}. The set of units of R is denoted $R^\times$.

NOTE The units of R form a group $R^\times$ under multiplication: 1 is a unit, the inverse of a unit is a unit, and if a, b are units then so is ab, with $(ab)^{-1} = b^{-1}a^{-1}$. For example $\mathbb{Z}^\times = \{1, -1\}$, and for $n \geqslant 2$, $(\mathbb{Z}/n)^\times$ is the group of residues coprime to n (Proposition 2.4.9). In a nonzero ring, 0 is never a unit, since $0b = 0 \neq 1$.

Definition 4.1.6 (Field)

A ***field*** is a commutative ring k in which $1 \neq 0$ and every nonzero element is a unit. Equivalently, $k^\times = k \setminus \{0\}$, so that $(k \setminus \{0\}, \cdot)$ is an abelian group.

The condition $1 \neq 0$ excludes the zero ring, which by convention is not a field. The rings $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ are fields; $\mathbb{Z}$ is not, since 2 has no inverse in $\mathbb{Z}$.

Definition 4.1.7 (Subrings and subfields)

A ***subring*** of a ring R is a subset $S \subseteq R$ that contains 0 and 1 and is closed under addition, negation and multiplication; it is a ring with the operations of R. A ***subfield*** of a field K is a subring $k \subseteq K$ that is closed under taking inverses of nonzero elements; it is then a field. In this situation we call K a ***field extension*** of k.

For instance $\mathbb{Z}$ is a subring of $\mathbb{Q}$, and $\mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}$ are subfields.

Definition 4.1.8 (Zero divisor)

A nonzero element a of a commutative ring R is a ***zero divisor*** if $ab = 0$ for some nonzero $b \in R$.

Proposition 4.1.9 (Fields have no zero divisors)

Let k be a field and $a, b \in k$. If $ab = 0$, then $a = 0$ or $b = 0$. Consequently, cancellation holds: if $a \neq 0$ and $ab = ac$, then $b = c$.

Proof. Suppose $ab = 0$ and $a \neq 0$. Then $b = 1b = (a^{-1}a)b = a^{-1}(ab) = a^{-1}0 = 0$. For cancellation, $ab = ac$ gives $a(b - c) = 0$, so $b - c = 0$. □

NOTE The converse fails: $\mathbb{Z}$ has no zero divisors and is not a field. A commutative ring with $1 \neq 0$ and no zero divisors is called an *integral domain*; these return in Chapter 17. In a general ring, products of nonzero elements can vanish, and then cancellation fails; Example 4.1.11 shows this happening in $\mathbb{Z}/6$.

4.1.1 The Rings $\mathbb{Z}/n$ and the Fields $\mathbb{F}_p$

Recall from Chapter 2 that $\mathbb{Z}/n = \{0, 1, \ldots, n-1\}$ is a group under addition mod n. We now also multiply mod n: the product of a and b in $\mathbb{Z}/n$ is the remainder of the integer ab on division by n.

Proposition 4.1.10 ($\mathbb{Z}/n$ is a ring)

For every $n \geqslant 1$, the set $\mathbb{Z}/n$ with addition and multiplication mod n is a commutative ring with identity 1. (For $n = 1$ it is the zero ring.)

Proof. Write $a \equiv a'$ if n divides $a - a'$. If $a \equiv a'$ and $b \equiv b'$, then $a + b \equiv a' + b'$ and $ab \equiv a'b'$, the latter because $ab - a'b' = a(b - b') + (a - a')b'$. Hence when we evaluate an expression built from sums and products, reducing mod n after every operation gives the same remainder as reducing only once at the end. For instance, for $a, b, c \in \mathbb{Z}/n$, both $(ab)c$ and $a(bc)$ computed in $\mathbb{Z}/n$ equal the remainder of the integer abc, so they are equal. Each of the other ring axioms is an identity that holds in $\mathbb{Z}$, and it holds in $\mathbb{Z}/n$ by the same argument. □

Example 4.1.11

In $\mathbb{Z}/6$ we have $2 \cdot 3 = 0$, so 2 and 3 are zero divisors and $\mathbb{Z}/6$ is not a field. Cancellation fails as well: $2 \cdot 1 = 2 \cdot 4$ in $\mathbb{Z}/6$, although $1 \neq 4$.

Proposition 4.1.12 ($\mathbb{Z}/p$ is a field)

Let $n \geqslant 2$. The ring $\mathbb{Z}/n$ is a field if and only if n is prime.

Proof. If n is not prime, write $n = ab$ with $1 < a, b < n$. Then a and b are nonzero in $\mathbb{Z}/n$ and $ab = 0$, so $\mathbb{Z}/n$ has zero divisors and is not a field.

Now let $n = p$ be prime. Since $p \geqslant 2$, we have $1 \neq 0$ in $\mathbb{Z}/p$. Let $a \in \mathbb{Z}/p$ be nonzero. The order of a in the additive group $\mathbb{Z}/p$ divides $|\mathbb{Z}/p| = p$ by Lagrange's theorem (Theorem 3.1.4), and it is not 1 because $a \neq 0$; so it equals p. Hence the p multiples $0, a, 2a, \ldots, (p-1)a$ in the additive group are distinct, and they exhaust $\mathbb{Z}/p$. In particular $\ell a = 1$ for some $\ell \in \{1, \ldots, p-1\}$. The ℓ-fold sum $a + \cdots + a$ equals $(1 + \cdots + 1)a$, the product of ℓ and a in the ring $\mathbb{Z}/p$, by distributivity. So ℓ is a multiplicative inverse of a. □

When we regard $\mathbb{Z}/p$ as a field we write it $\mathbb{F}_p$. (The proposition also follows from the description of $(\mathbb{Z}/n)^\times$ as the residues coprime to n in Proposition 2.4.9: for $n = p$ these are all the nonzero residues.)

4.2 Ring and Field Homomorphisms

Just as group homomorphisms preserve the group operation, ring and field homomorphisms preserve both operations.

Definition 4.2.1 (Ring homomorphism)

A ***ring homomorphism*** is a map $\varphi : R \to S$ between rings that respects both operations and the identity:

(a) $\varphi(a + b) = \varphi(a) + \varphi(b)$ for all $a, b \in R$;

(b) $\varphi(ab) = \varphi(a)\varphi(b)$ for all $a, b \in R$;

(c) $\varphi(1_R) = 1_S$.

A ***field homomorphism*** is a ring homomorphism between fields. A bijective ring homomorphism is an ***isomorphism***.

NOTE The first condition says that φ is a group homomorphism $(R, +) \to (S, +)$, so $\varphi(0) = 0$ and $\varphi(-a) = -\varphi(a)$ automatically. The third condition does *not* follow from the second, even for fields: the zero map $\varphi = 0$ satisfies $\varphi(ab) = 0 = \varphi(a)\varphi(b)$ but $\varphi(1) = 0 \neq 1$. So we include it as a separate axiom. The inverse of an isomorphism φ is again a ring homomorphism: writing $x = \varphi(a)$ and $y = \varphi(b)$, we get $\varphi^{-1}(xy) = \varphi^{-1}(\varphi(ab)) = ab = \varphi^{-1}(x)\varphi^{-1}(y)$, and similarly for sums.

Example 4.2.2

- Reduction mod n, the map $\mathbb{Z} \to \mathbb{Z}/n$ sending an integer to its remainder on division by n, is a ring homomorphism; this is the congruence property in the proof that $\mathbb{Z}/n$ is a ring.
- Complex conjugation $\mathbb{C} \to \mathbb{C}$, $z \mapsto \bar{z}$, is a field isomorphism, since $\overline{z + w} = \bar{z} + \bar{w}$, $\overline{zw} = \bar{z}\,\bar{w}$ and $\bar{1} = 1$.
- The inclusion of a subring into a ring is a ring homomorphism.

Proposition 4.2.3 (Field homomorphisms are injective)

If $\varphi : k \to K$ is a field homomorphism, then φ is injective.

Proof. Let $a \in k$ with $a \neq 0$, and let $b = a^{-1}$. Then

$$\varphi(a)\varphi(b) = \varphi(ab) = \varphi(1_k) = 1_K \neq 0_K,$$

using $1 \neq 0$ in the field K. If $\varphi(a)$ were 0, the left side would be 0; hence $\varphi(a) \neq 0$. So the kernel of the additive group homomorphism φ is $\{0\}$, and φ is injective (Proposition 2.5.6). □

NOTE So every field homomorphism $k \to K$ is an embedding: it identifies k with the subfield $\varphi(k)$ of K. Ring homomorphisms in general can collapse a great deal; reduction mod n, for instance, sends all of $n\mathbb{Z}$ to 0.

4.3 Polynomial Rings and Field Extensions

Polynomials connect algebra to analysis, number theory and geometry. Given a field k, we construct the ring of polynomials $k[x]$, some rings and fields built from it, and we see how to enlarge k so that it contains roots of polynomials.

4.3.1 The Polynomial Ring $k[x]$

Definition 4.3.1 (Polynomial ring)

Given a field k, the ***polynomial ring in one variable*** over k is

$$k[x] := \left\{ a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \;\middle|\; n \in \mathbb{N},\ a_i \in k \right\}.$$

Here x is a ***formal variable***: it is not an element of k or of any other set. Formally, a polynomial is a sequence $(a_0, a_1, a_2, \ldots)$ of elements of k with only finitely many nonzero terms. Polynomials are added coefficientwise, and multiplied by expanding and collecting powers of x using $x^i x^j = x^{i+j}$:

$$\Big(\sum_i a_i x^i\Big)\Big(\sum_j b_j x^j\Big) = \sum_n c_n x^n, \qquad c_n = \sum_{i+j=n} a_i b_j.$$

Multiplication is *not* coefficientwise: $(1+x)(1+x) = 1 + 2x + x^2$. One checks from these formulas that $k[x]$ is a commutative ring with identity the constant polynomial 1. The only axiom that needs a computation is associativity of multiplication: for $f = \sum a_i x^i$, $g = \sum b_j x^j$ and $h = \sum c_l x^l$, the coefficient of x^n in both $(fg)h$ and $f(gh)$ is $\sum_{i+j+l=n} a_i b_j c_l$. The constants form a subring of $k[x]$ that we identify with k.

NOTE The adjective "formal" matters. When we write $x^2 + 1 \in \mathbb{R}[x]$, we are not thinking of x as a real number; $x^2 + 1$ is a symbolic expression. However, we *can* evaluate a polynomial at any element of k, or of any

field K containing k: given $f = a_0 + a_1 x + \cdots + a_n x^n$ and $r \in K$, we set $f(r) = a_0 + a_1 r + \cdots + a_n r^n \in K$. Comparing with the product formula shows $(f+g)(r) = f(r) + g(r)$ and $(fg)(r) = f(r)g(r)$, so evaluation at r is a ring homomorphism $k[x] \to K$.

Definition 4.3.2 (Degree)

The ***degree*** $\deg f$ of a nonzero polynomial $f = \sum a_i x^i$ is the largest n with $a_n \neq 0$, and a_n is its ***leading coefficient***. The polynomial is ***monic*** if its leading coefficient is 1. Nonzero constants have degree 0; by convention $\deg 0 = -\infty$.

Lemma 4.3.3 (Degree of a product)

If $f, g \in k[x]$ are nonzero, then $fg \neq 0$ and $\deg(fg) = \deg f + \deg g$. In particular, $k[x]$ has no zero divisors.

Proof. Write $f = \sum_{i \leqslant m} a_i x^i$ and $g = \sum_{j \leqslant n} b_j x^j$ with $a_m \neq 0$ and $b_n \neq 0$. In the coefficient $c_l = \sum_{i+j=l} a_i b_j$ of fg, every term with $i > m$ or $j > n$ vanishes. If $l > m + n$, every pair with $i + j = l$ has $i > m$ or $j > n$, so $c_l = 0$. For $l = m + n$, the only surviving term is $c_{m+n} = a_m b_n$, which is nonzero because k has no zero divisors. □

Proposition 4.3.4 (Units of $k[x]$)

The units of $k[x]$ are exactly the nonzero constants. In particular, $k[x]$ is not a field.

Proof. A nonzero constant c has inverse c^{-1}. Conversely, if $fg = 1$, then $f, g \neq 0$ and $\deg f + \deg g = \deg 1 = 0$ by the lemma. Both degrees are nonnegative, so $\deg f = 0$ and f is a nonzero constant. For instance, x has no inverse in $k[x]$. □

4.3.2 The Field of Rational Functions

To make $k[x]$ into a field we allow fractions of polynomials, just as we pass from $\mathbb{Z}$ to $\mathbb{Q}$.

Definition 4.3.5 (Field of rational functions)

The ***field of rational functions*** over k is

$$k(x) := \left\{ \frac{p}{q} \;\middle|\; p, q \in k[x],\ q \neq 0 \right\} \Big/ \sim, \qquad \text{where } \frac{p}{q} \sim \frac{p'}{q'} \text{ iff } pq' = p'q.$$

Addition and multiplication are defined as for ordinary fractions:

$$\frac{p}{q} + \frac{p'}{q'} = \frac{pq' + p'q}{qq'}, \qquad \frac{p}{q} \cdot \frac{p'}{q'} = \frac{pp'}{qq'}.$$

The denominators qq' are nonzero by Lemma 4.3.3. As for $\mathbb{Q}$, one checks that $\sim$ is an equivalence relation (transitivity uses cancellation in $k[x]$, which holds because $k[x]$ has no zero divisors), that the operations do not depend on the chosen representatives, and that the ring axioms hold. The inverse of a nonzero element p/q is q/p, so $k(x)$ is a field. Identifying p with $p/1$ makes $k[x]$ a subring of $k(x)$.

This construction generalizes to polynomials in any number of variables, giving the field $k(x_1, \ldots, x_n)$ of rational functions in n variables.

4.3.3 Formal Power Series

Another way to extend $k[x]$ is to allow *infinite* sums.

Definition 4.3.6 (Formal power series)

The ***ring of formal power series*** over k is

$$k[\![x]\!] := \left\{ \sum_{i=0}^{\infty} a_i x^i \;\middle|\; a_i \in k \right\}.$$

Addition is coefficientwise, and multiplication is given by the same formula as for polynomials: for $f = \sum a_i x^i$ and $g = \sum b_i x^i$, we set $fg = \sum c_n x^n$ with $c_n = \sum_{i+j=n} a_i b_j$.

NOTE The product formula makes sense because each coefficient c_n is a *finite* sum: only the $n+1$ pairs $(i, j) = (0, n), \ldots, (n, 0)$ satisfy $i + j = n$ with $i, j \geqslant 0$. This is why these power series are called "formal": no question of convergence arises. The ring axioms are checked as for $k[x]$, and $k[\![x]\!]$

contains $k[x]$ as a subring (polynomials are the power series with only finitely many nonzero coefficients).

Unlike $k[x]$, the ring $k[\![x]\!]$ has many units.

Lemma 4.3.7 (Units of $k[\![x]\!]$)
A power series $f = \sum_{i=0}^{\infty} a_i x^i \in k[\![x]\!]$ has a multiplicative inverse if and only if $a_0 \neq 0$.

Proof. ($\Rightarrow$) If $fg = 1$ for some $g = \sum b_i x^i$, then comparing constant terms gives $a_0 b_0 = 1$, so $a_0 \neq 0$.

($\Leftarrow$) Assume $a_0 \neq 0$. We construct $g = \sum b_i x^i$ with $fg = 1$ one coefficient at a time. The constant term of fg must be $a_0 b_0 = 1$, so we take $b_0 = a_0^{-1}$. For $n \geqslant 1$, the coefficient of x^n in fg must vanish:

$$a_0 b_n + \sum_{i=1}^{n} a_i b_{n-i} = 0, \qquad \text{i.e.} \qquad b_n = -a_0^{-1} \sum_{i=1}^{n} a_i b_{n-i}.$$

The right side involves only $b_0, \ldots, b_{n-1}$, so these equations determine $b_1, b_2, \ldots$ inductively, and the resulting g satisfies $fg = 1$. □

Example 4.3.8
The geometric series: in $k[\![x]\!]$ we have $(1-x)^{-1} = 1+x+x^2+x^3+\cdots = \sum_{i=0}^{\infty} x^i$. Indeed, the coefficient of x^n in $(1-x)\sum_{i \geqslant 0} x^i$ is $1-1 = 0$ for $n \geqslant 1$, and the constant term is 1.

4.3.4 Laurent Series

Every nonzero element of $k[\![x]\!]$ can be written as $x^m(a_m + a_{m+1}x + \cdots)$, where a_m is its first nonzero coefficient. The second factor is a unit by Lemma 4.3.7, so to obtain a field we need only invert x, that is, allow finitely many negative powers of x.

Definition 4.3.9 (Field of Laurent series)

The ***field of Laurent series*** over k is

$$k(\!(x)\!) := \left\{ \sum_{i=m}^{\infty} a_i x^i \;\middle|\; m \in \mathbb{Z},\ a_i \in k \right\},$$

the set of formal series in which only finitely many negative powers of x appear. Addition is coefficientwise and multiplication is again $\left(\sum a_i x^i\right)\left(\sum b_j x^j\right) = \sum c_n x^n$ with $c_n = \sum_{i+j=n} a_i b_j$.

If $a_i = 0$ for $i < m$ and $b_j = 0$ for $j < m'$, the sum defining c_n runs only over $m \leqslant i \leqslant n - m'$, a finite range, and $c_n = 0$ for $n < m + m'$; so the product is again a Laurent series. As for $k[\![x]\!]$, this makes $k((x))$ a commutative ring containing $k[\![x]\!]$ as a subring.

Proposition 4.3.10 ($k((x))$ is a field)

Every nonzero Laurent series has a multiplicative inverse.

Proof. Let $f = \sum_{i=m}^{\infty} a_i x^i$ with $a_m \neq 0$. Then $f = x^m g$, where $g = a_m + a_{m+1}x + \cdots \in k[\![x]\!]$ has nonzero constant term. By Lemma 4.3.7, g has an inverse g^{-1} in $k[\![x]\!]$. Then $f^{-1} = x^{-m} g^{-1} \in k((x))$. □

4.3.5 Roots of Polynomials and Field Extensions

Given a polynomial $f \in k[x]$ with $\deg f \geqslant 1$, we can ask whether f has a root in k, and if not, whether we can *extend* k to a larger field containing one.

Definition 4.3.11 (Root)

Let K be a field extension of k and $f \in k[x]$. An element $r \in K$ is a ***root*** of f if $f(r) = 0$. We write K/k for the extension and sometimes $k \subseteq K$.

Example 4.3.12

The polynomial $x^2 - 2 \in \mathbb{Q}[x]$ has no roots in $\mathbb{Q}$, since $\sqrt{2}$ is irrational. Consider

$$\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} : a, b \in \mathbb{Q}\} \subseteq \mathbb{R}.$$

It contains $\mathbb{Q}$ and the root $\sqrt{2}$ of $x^2 - 2$, and it is closed under addition and under multiplication, since $(a + b\sqrt{2})(c + d\sqrt{2}) = (ac + 2bd) + (ad + bc)\sqrt{2}$. It is also closed under inverses: for $a + b\sqrt{2} \neq 0$,

$$\frac{1}{a + b\sqrt{2}} = \frac{a - b\sqrt{2}}{a^2 - 2b^2} = \frac{a}{a^2 - 2b^2} - \frac{b}{a^2 - 2b^2}\sqrt{2} \in \mathbb{Q}(\sqrt{2}).$$

Here $a^2 - 2b^2 \neq 0$: if $b \neq 0$, then $a^2 = 2b^2$ would give $\sqrt{2} = |a/b| \in \mathbb{Q}$, and if $b = 0$, then $a \neq 0$. So $\mathbb{Q}(\sqrt{2})$ is a subfield of $\mathbb{R}$, and a field extension of $\mathbb{Q}$ containing a root of $x^2 - 2$.

Example 4.3.13

Similarly, $x^2 + 1 \in \mathbb{R}[x]$ has no real roots. The extension

$$\mathbb{R}(i) = \mathbb{R}(\sqrt{-1}) = \{a + bi : a, b \in \mathbb{R}\} = \mathbb{C}$$

is the field of complex numbers, which contains the roots $\pm i$ of $x^2 + 1$.

Lemma 4.3.14 (Factor theorem)

Let $f \in k[x]$ have degree $n \geqslant 1$ and let $r \in k$. Then $f = (x - r)q + f(r)$ for some $q \in k[x]$ of degree $n - 1$. In particular, r is a root of f if and only if $f = (x - r)q$ for some $q \in k[x]$.

Proof. For $i \geqslant 1$ we have $x^i - r^i = (x-r)(x^{i-1} + rx^{i-2} + \cdots + r^{i-1})$. If $f = \sum_{i=0}^{n} a_i x^i$, then

$$f - f(r) = \sum_{i=1}^{n} a_i(x^i - r^i) = (x - r)q, \qquad q = \sum_{i=1}^{n} a_i(x^{i-1} + rx^{i-2} + \cdots + r^{i-1}),$$

and the coefficient of x^{n-1} in q is $a_n \neq 0$. If $f(r) = 0$ this gives $f = (x - r)q$. Conversely, if $f = (x - r)q$, evaluating at r gives $f(r) = 0 \cdot q(r) = 0$. □

Corollary 4.3.15 (Number of roots)

A nonzero polynomial $f \in k[x]$ of degree n has at most n roots in k.

Proof. By induction on n. A nonzero constant has no roots. Let $n \geqslant 1$. If f has no root we are done. Otherwise let r be a root and write $f = (x - r)q$ with $\deg q = n - 1$. If $s \neq r$ is another root, then $0 = f(s) = (s - r)q(s)$ with $s - r \neq 0$, so $q(s) = 0$. Thus every root of f other than r is a root of q, and q has at most $n - 1$ roots by induction. □

Definition 4.3.16 (Algebraically closed field)

A field k is ***algebraically closed*** if every non-constant polynomial $f \in k[x]$ has a root in k.

Proposition 4.3.17 (Factorization into linear factors)

A field k is algebraically closed if and only if every polynomial $f \in k[x]$ of degree $n \geqslant 1$ factors as

$$f = c(x - r_1)(x - r_2)\cdots(x - r_n), \qquad c \in k \setminus \{0\},\ r_1, \ldots, r_n \in k$$

(the r_i need not be distinct).

Proof. If every such f factors in this way, then r_1 is a root of f, so k is algebraically closed. Conversely, let k be algebraically closed; we induct on n. If $n = 1$, then $f = a_1 x + a_0 = a_1\big(x - (-a_0 a_1^{-1})\big)$. If $n > 1$, let r_1 be a root of f. By the factor theorem, $f = (x - r_1)q$ with $\deg q = n - 1 \geqslant 1$, and by induction $q = c(x - r_2)\cdots(x - r_n)$. □

NOTE The Fundamental Theorem of Algebra states that $\mathbb{C}$ is algebraically closed; we do not prove it here. Over an algebraically closed field, every non-constant polynomial already has a root, so adjoining roots of polynomials produces no larger fields. The field $\mathbb{C}$ is called the *algebraic closure* of $\mathbb{R}$: it is algebraically closed, and every element of $\mathbb{C}$ is a root of a nonzero real polynomial ($a + bi$ is a root of $x^2 - 2ax + a^2 + b^2$). By contrast, $\mathbb{Q}$ is very far from algebraically closed: $x^2 - 2$, $x^2 - 3$, $x^3 - 2$, … have no rational roots, and adjoining their roots gives infinitely many different extensions of $\mathbb{Q}$.

4.4 The Characteristic of a Field

Every ring receives a distinguished homomorphism from the integers. For a field, its kernel is an important invariant.

For an element a of a ring and $n \in \mathbb{Z}$, we write $n \cdot a$ for the n-th multiple of a in the additive group: $n \cdot a = a + \cdots + a$ (n terms) for $n > 0$, $0 \cdot a = 0$, and $n \cdot a = -((-n) \cdot a)$ for $n < 0$. This is the additive version of the powers a^n in a group from Chapter 2, so $(m + n) \cdot a = m \cdot a + n \cdot a$ and $m \cdot (n \cdot a) = (mn) \cdot a$. Distributivity and Proposition 4.1.3 give $(n \cdot a)b = n \cdot (ab)$.

Proposition 4.4.1 (The homomorphism from $\mathbb{Z}$)

For any ring R there is a unique ring homomorphism $\varphi : \mathbb{Z} \to R$, namely $\varphi(n) = n \cdot 1_R$.

Proof. Uniqueness: a ring homomorphism φ is in particular a homomorphism of additive groups with $\varphi(1) = 1_R$. Since 1 generates $(\mathbb{Z}, +)$, it is determined by this value: $\varphi(n) = n \cdot \varphi(1) = n \cdot 1_R$.

Existence: define $\varphi(n) = n \cdot 1_R$. Then $\varphi(m+n) = \varphi(m)+\varphi(n)$ and $\varphi(1) = 1_R$. Finally,

$$\varphi(m)\varphi(n) = (m \cdot 1_R)(n \cdot 1_R) = m \cdot \big(1_R(n \cdot 1_R)\big) = m \cdot (n \cdot 1_R) = (mn) \cdot 1_R = \varphi(mn).$$

□

Now let k be a field and $\varphi : \mathbb{Z} \to k$ this homomorphism. Its kernel is a subgroup of $\mathbb{Z}$, hence equal to $n\mathbb{Z}$ for a unique $n \geqslant 0$ (Theorem 2.4.5).

Proposition 4.4.2 (The kernel is zero or prime)
The kernel of $\varphi : \mathbb{Z} \to k$ is $n\mathbb{Z}$ where either $n = 0$ or $n = p$ is prime.

Proof. We have $n \neq 1$, since $\varphi(1) = 1_k \neq 0$. Suppose $n > 1$ is not prime, so $n = ab$ with $1 < a, b < n$. Then

$$\varphi(a)\varphi(b) = \varphi(ab) = \varphi(n) = 0,$$

and since a field has no zero divisors, $\varphi(a) = 0$ or $\varphi(b) = 0$. That is, $a \in n\mathbb{Z}$ or $b \in n\mathbb{Z}$, which is impossible for $0 < a, b < n$. Hence if $n > 0$, then n is prime. □

Definition 4.4.3 (Characteristic)

The ***characteristic*** of a field k, denoted $\operatorname{char}(k)$, is

- 0 if $\ker \varphi = \{0\}$, i.e., $\varphi : \mathbb{Z} \to k$ is injective;
- p if $\ker \varphi = p\mathbb{Z}$ for a prime p.

Equivalently, $\operatorname{char}(k) = p > 0$ means that $\underbrace{1_k + 1_k + \cdots + 1_k}_{p \text{ terms}} = 0_k$ and p is the smallest positive integer with this property, while $\operatorname{char}(k) = 0$ means that $n \cdot 1_k \neq 0$ for every $n \geqslant 1$.

In a field of characteristic p, we have $p \cdot a = (p \cdot 1)a = 0$ for every a. In characteristic 2 in particular, $1 + 1 = 0$, that is, $-1 = 1$.

Example 4.4.4

- $\operatorname{char}(\mathbb{Q}) = \operatorname{char}(\mathbb{R}) = \operatorname{char}(\mathbb{C}) = 0$: the map $\mathbb{Z} \to \mathbb{Q}$ is the inclusion.
- $\operatorname{char}(\mathbb{F}_p) = p$ for every prime p: $p \cdot 1 = 0$ in $\mathbb{Z}/p$, while $n \cdot 1 = n \neq 0$ for $0 < n < p$.

- $\operatorname{char}(k(x)) = \operatorname{char}(k((x))) = \operatorname{char}(k)$: these fields contain k as a subfield, so the homomorphism from $\mathbb{Z}$ to either of them is $\mathbb{Z} \to k$ followed by the inclusion, and has the same kernel.

NOTE The field $\mathbb{F}_p = \mathbb{Z}/p$ is the ***prime field*** of characteristic p. Every field k of characteristic p contains a copy of $\mathbb{F}_p$: since $a \cdot 1_k = b \cdot 1_k$ exactly when p divides $a - b$, the map $\mathbb{F}_p \to k$, $a \mapsto a \cdot 1_k$, is an injective ring homomorphism. Every field of characteristic 0 contains a copy of $\mathbb{Q}$, via $m/n \mapsto (m \cdot 1_k)(n \cdot 1_k)^{-1}$. Thus $\mathbb{F}_p$ and $\mathbb{Q}$ are the smallest fields of their characteristics.

Theorem 4.4.5 (Existence and uniqueness of finite fields)

For every prime p and every $n \geqslant 1$, there exists a field with exactly p^n elements, unique up to isomorphism, denoted $\mathbb{F}_{p^n}$ or $\mathrm{GF}(p^n)$. Every finite field is of this form.

NOTE We do not prove this theorem here. The field $\mathbb{F}_{p^n}$ can be constructed as a splitting field of $x^{p^n} - x$ over $\mathbb{F}_p$ (splitting fields are discussed in Chapter 17). There are also infinite fields of characteristic p, for example the field of rational functions $\mathbb{F}_p(x)$ and the field of Laurent series $\mathbb{F}_p((x))$.

4.5 Vector Spaces

We now arrive at the central structure of linear algebra. A vector space is a set with two operations, addition and scalar multiplication, satisfying axioms modelled on the familiar properties of $\mathbb{R}^n$. Throughout, k denotes a field.

Definition 4.5.1 (Vector space)

Fix a field k. A ***vector space over*** k (or k***-vector space***) is a set V equipped with two operations, **vector addition** $+ : V \times V \to V$ and **scalar multiplication** $k \times V \to V$, $(a, v) \mapsto a \cdot v = av$, satisfying the following axioms:

(a) $(V, +)$ is an abelian group, whose identity 0 is called the ***zero vector***;

(b) $1 \cdot v = v$ for all $v \in V$;

(c) $(ab) \cdot v = a \cdot (b \cdot v)$ for all $a, b \in k$ and $v \in V$;

(d) $(a + b) \cdot v = a \cdot v + b \cdot v$ for all $a, b \in k$ and $v \in V$;

(e) $a \cdot (v + w) = a \cdot v + a \cdot w$ for all $a \in k$ and $v, w \in V$.

Elements of V are called ***vectors***; elements of k are called ***scalars***.

Axioms (b) and (c) express that 1 acts as the identity and that scalar multiplication is associative; the two distributive laws (d) and (e) connect scalar multiplication with the two additions.

Proposition 4.5.2 (First consequences of the axioms)

Let V be a vector space over k, $a \in k$ and $v \in V$. Then:

(a) $0 \cdot v = 0$ and $a \cdot 0 = 0$;

(b) $(-1) \cdot v = -v$;

(c) if $a \cdot v = 0$, then $a = 0$ or $v = 0$.

Proof. (a) We have $0 \cdot v = (0 + 0) \cdot v = 0 \cdot v + 0 \cdot v$; adding $-(0 \cdot v)$ to both sides gives $0 \cdot v = 0$. Similarly $a \cdot 0 = a \cdot (0 + 0) = a \cdot 0 + a \cdot 0$, so $a \cdot 0 = 0$.

(b) $v + (-1) \cdot v = 1 \cdot v + (-1) \cdot v = (1 + (-1)) \cdot v = 0 \cdot v = 0$.

(c) If $a \cdot v = 0$ and $a \neq 0$, then $v = 1 \cdot v = (a^{-1}a) \cdot v = a^{-1} \cdot (a \cdot v) = a^{-1} \cdot 0 = 0$. □

4.5.1 Fundamental Examples

Example 4.5.3

The standard vector space k^n: For $n \geqslant 1$, the set

$$k^n = \{(a_1, a_2, \ldots, a_n) : a_i \in k\}$$

is a vector space with componentwise operations:

$$(a_1, \ldots, a_n) + (b_1, \ldots, b_n) = (a_1 + b_1, \ldots, a_n + b_n),$$
$$c \cdot (a_1, \ldots, a_n) = (ca_1, \ldots, ca_n).$$

We also set $k^0 = \{0\}$. This is the prototypical n-dimensional vector space.

Example 4.5.4

Infinite sequences k^∞: The set of all sequences in k,

$$k^\infty = \{(a_i)_{i \in \mathbb{N}} : a_i \in k\} = \{(a_0, a_1, a_2, \ldots) : a_i \in k\},$$

is a vector space with componentwise operations. It contains the ***eventually zero sequences***, those with $a_i = 0$ for all sufficiently large i, which form a vector space under the same operations.

Example 4.5.5

Function spaces: For any set S, the set

$$k^S = \{f : S \to k\}$$

of all functions from S to k is a vector space with pointwise operations: $(f+g)(s) = f(s) + g(s)$ and $(c \cdot f)(s) = c \cdot f(s)$. When $S = \mathbb{N}$ we recover k^∞, and when $S = \{1, \ldots, n\}$ we recover k^n.

Example 4.5.6

Polynomials and power series:

- $k[x]$ is a k-vector space: addition is polynomial addition, and scalar multiplication multiplies each coefficient by the scalar.
- $k[\![x]\!]$ is a k-vector space in the same way.
- For each $d \geqslant 0$, the set $k[x]_{\leqslant d} = \{f \in k[x] : \deg f \leqslant d\}$ of polynomials of degree at most d is a k-vector space.

Example 4.5.7

Spaces of functions $\mathbb{R} \to \mathbb{R}$: The space $\mathbb{R}^\mathbb{R}$ of all functions $\mathbb{R} \to \mathbb{R}$ contains many interesting vector spaces, each closed under addition and scalar multiplication:

- the continuous functions $C(\mathbb{R})$;
- the differentiable functions;
- the continuously differentiable functions $C^1(\mathbb{R})$;
- the smooth (infinitely differentiable) functions $C^\infty(\mathbb{R})$;
- the polynomial functions $r \mapsto a_0 + a_1 r + \cdots + a_n r^n$.

Each is contained in the one before it. A polynomial $f \in \mathbb{R}[x]$ defines the polynomial function $r \mapsto f(r)$, and two different polynomials define different functions: their difference is a nonzero polynomial, which by Corollary 4.3.15 has only finitely many roots, while $\mathbb{R}$ is infinite. So we may identify $\mathbb{R}[x]$ with the space of real polynomial functions. Over a finite field this identification

fails: over $\mathbb{F}_p$, the polynomials x^p and x define the same function $\mathbb{F}_p \to \mathbb{F}_p$, because $a^{p-1} = 1$ for every a in the group $\mathbb{F}_p^\times$ of order $p-1$ (a corollary of Lagrange's theorem).

4.5.2 Subspaces

Definition 4.5.8 (Subspace)

A ***subspace*** of a vector space V is a nonempty subset $W \subseteq V$ that is closed under addition and scalar multiplication:

(a) $w_1, w_2 \in W \implies w_1 + w_2 \in W$;

(b) $w \in W,\ a \in k \implies a \cdot w \in W$.

A subspace is a vector space under the operations inherited from V.

Proposition 4.5.9 (Subspace criterion)

A nonempty subset $W \subseteq V$ is a subspace if and only if $aw_1 + bw_2 \in W$ for all $a, b \in k$ and $w_1, w_2 \in W$.

Proof. ($\Rightarrow$) If W is a subspace, then $aw_1 \in W$ and $bw_2 \in W$ by closure under scalar multiplication, and $aw_1 + bw_2 \in W$ by closure under addition.

($\Leftarrow$) Taking $a = b = 1$ gives closure under addition. Taking $b = 0$ gives closure under scalar multiplication. □

NOTE A subspace always contains 0: if $w \in W$, then $0 = 0 \cdot w \in W$ by Proposition 4.5.2. So in the definition we may replace "nonempty" by "contains 0". Some condition of this kind is needed, since the empty set is closed under both operations but is not a vector space. With it, a subspace is precisely a subset that contains 0 and is closed under linear combinations, and it contains $-w = (-1) \cdot w$ along with each w.

Example 4.5.10

- $\{0\}$ and V itself are always subspaces of V.
- In $\mathbb{R}^3$, every line through the origin and every plane through the origin is a subspace. Once we have the notion of dimension (Section 4.7), it will follow that, together with $\{0\}$ and $\mathbb{R}^3$, these are the only subspaces of $\mathbb{R}^3$.

- The eventually zero sequences form a subspace of k^{∞}, and $k[x]$ is a subspace of $k[\![x]\!]$.
- In $k[x]$, the polynomials of degree at most n form a subspace. The polynomials of degree *exactly* n do *not*: the zero polynomial is missing, and for $n \geqslant 1$ the set is not closed under addition, since $x^n + (1 - x^n) = 1$.

Proposition 4.5.11 (Intersections of subspaces)

The intersection of any nonempty family of subspaces of V is a subspace of V.

Proof. Let $W = \bigcap_{i \in I} W_i$. Each W_i contains 0, so $0 \in W$. If $w_1, w_2 \in W$ and $a, b \in k$, then $aw_1 + bw_2 \in W_i$ for every i, so $aw_1 + bw_2 \in W$. □

4.5.3 Linear Maps and Isomorphisms

To compare vector spaces we need maps that respect their structure. Chapter 5 studies these maps in detail; here we introduce them, since we need the notion of isomorphism right away.

Definition 4.5.12 (Linear map)

Let V and W be vector spaces over k. A ***linear map*** (or ***k-linear map***, or ***homomorphism of vector spaces***) is a function $\varphi : V \to W$ such that

(a) $\varphi(u + v) = \varphi(u) + \varphi(v)$ for all $u, v \in V$;

(b) $\varphi(\lambda v) = \lambda \varphi(v)$ for all $\lambda \in k$ and $v \in V$.

A bijective linear map is an ***isomorphism***. We say V and W are ***isomorphic***, written $V \cong W$, if there is an isomorphism $V \to W$.

By induction on n, a linear map preserves all linear combinations:

$$\varphi\left(\sum_{i=1}^{n} a_i v_i\right) = \sum_{i=1}^{n} a_i \varphi(v_i).$$

In particular $\varphi(0) = 0$.

Proposition 4.5.13 (Composites and inverses)

The composite of linear maps is linear, and the inverse of an isomorphism is linear. Consequently, $\cong$ is an equivalence relation.

Proof. If $\varphi : U \to V$ and $\psi : V \to W$ are linear, then $\psi(\varphi(au + bv)) = \psi(a\varphi(u) + b\varphi(v)) = a\,\psi(\varphi(u)) + b\,\psi(\varphi(v))$. If $\varphi : V \to W$ is a linear bijection, $w, w' \in W$ and $a, b \in k$, then $\varphi\big(a\varphi^{-1}(w) + b\varphi^{-1}(w')\big) = aw + bw'$, so $\varphi^{-1}(aw + bw') = a\varphi^{-1}(w) + b\varphi^{-1}(w')$. Hence $\cong$ is reflexive (via id_V), symmetric (via inverses) and transitive (via composites). □

Example 4.5.14

$k[\![x]\!] \cong k^\infty$**:** The map

$$k[\![x]\!] \to k^\infty, \qquad \sum_{i=0}^{\infty} a_i x^i \mapsto (a_0, a_1, a_2, \ldots)$$

is bijective, since a power series is nothing but its sequence of coefficients, and it is linear, since both spaces add and scale coefficientwise. It is therefore an isomorphism of vector spaces. It restricts to an isomorphism from $k[x]$ onto the subspace of eventually zero sequences. The isomorphism ignores the multiplication of power series, which is not coefficientwise: as vector spaces the two are the same, as rings they are not.

4.5.4 Sums and Direct Sums of Subspaces

When can we decompose a vector space into independent pieces? The notion of direct sum captures exactly this.

Definition 4.5.15 (Sum of subspaces)

Given subspaces $U, W \subseteq V$, their ***sum*** is

$$U + W = \{u + w : u \in U,\ w \in W\}.$$

More generally, the sum of subspaces $U_1, \ldots, U_n$ is $U_1 + \cdots + U_n = \{u_1 + \cdots + u_n : u_i \in U_i\}$.

Proposition 4.5.16 (Sum is a subspace)

For subspaces $U, W \subseteq V$, the sum $U + W$ is a subspace of V, and it is the smallest subspace of V containing both U and W.

Proof. The zero vector $0 = 0 + 0$ lies in $U + W$. Let $v_1 = u_1 + w_1$ and $v_2 = u_2 + w_2$ be elements of $U + W$, and let $\lambda \in k$. Then

$$v_1 + v_2 = (u_1 + u_2) + (w_1 + w_2) \in U + W$$

since $u_1 + u_2 \in U$ and $w_1 + w_2 \in W$. Similarly, $\lambda v_1 = \lambda u_1 + \lambda w_1 \in U + W$. It contains U (take $w = 0$) and W (take $u = 0$), and any subspace containing U and W contains all sums $u + w$. □

The same argument shows that $U_1 + \cdots + U_n$ is the smallest subspace containing $U_1, \ldots, U_n$.

Definition 4.5.17 (Direct sum of two subspaces)

The sum $U + W$ is called ***direct***, and written $U \oplus W$, if $U \cap W = \{0\}$.

Proposition 4.5.18 (Characterizations of direct sums)

For subspaces $U, W \subseteq V$, the following are equivalent:

(a) $U + W$ is direct, i.e., $U \cap W = \{0\}$;

(b) every $v \in U + W$ has a unique expression $v = u + w$ with $u \in U$, $w \in W$;

(c) if $u + w = 0$ with $u \in U$ and $w \in W$, then $u = w = 0$.

Proof. (a) $\Rightarrow$ (b): Suppose $v = u_1 + w_1 = u_2 + w_2$. Then $u_1 - u_2 = w_2 - w_1$. The left side is in U and the right side is in W, so both lie in $U \cap W = \{0\}$. Thus $u_1 = u_2$ and $w_1 = w_2$.

(b) $\Rightarrow$ (c): If $u + w = 0$, then $u + w = 0 + 0$, and uniqueness gives $u = 0$ and $w = 0$.

(c) $\Rightarrow$ (a): If $v \in U \cap W$, then $v + (-v) = 0$ with $v \in U$ and $-v \in W$. By (c), $v = 0$. □

Definition 4.5.19 (Direct sum decomposition)

Subspaces $U_1, \ldots, U_n$ of V are ***independent*** if $u_1 + \cdots + u_n = 0$ with $u_i \in U_i$ implies $u_i = 0$ for all i. We say V is the ***(internal) direct sum*** of $U_1, \ldots, U_n$, written

$$V = U_1 \oplus U_2 \oplus \cdots \oplus U_n,$$

if the U_i are independent and $V = U_1 + \cdots + U_n$. Equivalently, every $v \in V$ can be written *uniquely* as $v = u_1 + \cdots + u_n$ with $u_i \in U_i$.

The equivalence is proved as in the case of two subspaces: two expressions $\sum u_i = \sum u_i'$ give $\sum (u_i - u_i') = 0$, and conversely uniqueness of the expression for 0 is independence. For two subspaces, $V = U \oplus W$ means $V = U + W$ and $U \cap W = \{0\}$.

NOTE For three or more subspaces, pairwise trivial intersections do *not* imply independence. In k^2, the three lines $U_1 = \{(a, 0)\}$, $U_2 = \{(0, a)\}$ and $U_3 = \{(a, a)\}$ meet pairwise only in 0, but $(1, 0) + (0, 1) + (-1, -1) = 0$.

Example 4.5.20

- In $\mathbb{R}^3$, let $U = \{(x, 0, 0) : x \in \mathbb{R}\}$ (the x-axis) and $W = \{(0, y, z) : y, z \in \mathbb{R}\}$ (the yz-plane). Then $\mathbb{R}^3 = U \oplus W$.
- In k^n, fix $0 \leqslant m \leqslant n$, let U consist of the vectors whose last $n - m$ entries are 0, and let W consist of those whose first m entries are 0. Then $k^n = U \oplus W$.
- Let $\mathcal{M}_n(k)$ be the vector space of $n \times n$ arrays $M = (m_{ij})$ of elements of k with entrywise operations (that is, k^S for $S = \{1, \ldots, n\}^2$; we study such matrices in Chapter 5), and let $M^t = (m_{ji})$ be the transpose of M. Suppose $\operatorname{char}(k) \neq 2$. The symmetric matrices $S = \{M : M^t = M\}$ and the antisymmetric matrices $A = \{M : M^t = -M\}$ are subspaces with $\mathcal{M}_n(k) = S \oplus A$. Indeed, every M is $\frac{1}{2}(M + M^t) + \frac{1}{2}(M - M^t)$ with the first term in S and the second in A, and if $M \in S \cap A$ then $M = -M$, so $2M = 0$ and $M = 0$. This fails in characteristic 2, where $\frac{1}{2}$ does not exist and $-M = M$, so that $A = S$.

NOTE Direct sums appear throughout linear algebra. Diagonalizing a linear operator means decomposing the space as a direct sum of eigenspaces, and if $V = U \oplus W$ then the quotient space V/U is isomorphic to W; both ideas are developed in Chapter 7. A direct sum decomposition lets us break a problem into independent pieces.

4.5.5 External Direct Sums and Products

We can also build new vector spaces from old ones by taking products.

Definition 4.5.21 (Direct product (external direct sum))

Given vector spaces $V_1, \ldots, V_n$ over the same field k, their ***direct product*** (or ***external direct sum***) is

$$V_1 \times V_2 \times \cdots \times V_n = \{(v_1, v_2, \ldots, v_n) : v_i \in V_i\}$$

with componentwise operations:

$$(v_1,\dots,v_n)+(w_1,\dots,w_n)=(v_1+w_1,\dots,v_n+w_n),$$
$$\lambda(v_1,\dots,v_n)=(\lambda v_1,\dots,\lambda v_n).$$

It is also denoted $V_1\oplus V_2\oplus\cdots\oplus V_n$ or $\bigoplus_{i=1}^n V_i$.

NOTE The external and internal direct sums are related as follows. If $V=U\oplus W$, then the map

$$U\times W\to V,\quad (u,w)\mapsto u+w,$$

is linear, surjective because $V=U+W$, and injective because expressions $u+w$ are unique; so it is an isomorphism. Conversely, $V_1\times V_2$ contains the subspaces $V_1\times\{0\}\cong V_1$ and $\{0\}\times V_2\cong V_2$, and it is their internal direct sum. The same holds for n summands.

Example 4.5.22

The space $\mathbb{R}^{m+n}$ is isomorphic to $\mathbb{R}^m\times\mathbb{R}^n$ via $(x_1,\dots,x_m,y_1,\dots,y_n)\mapsto((x_1,\dots,x_m),(y_1,\dots,y_n))$. This is why we freely identify $\mathbb{R}^m\oplus\mathbb{R}^n$ with $\mathbb{R}^{m+n}$.

For an infinite family of vector spaces $(V_i)_{i\in I}$, the two constructions differ.

Definition 4.5.23 (Infinite direct sums and products)

Let $(V_i)_{i\in I}$ be a family of vector spaces over k. Their ***direct product*** and ***direct sum*** are

$$\prod_{i\in I}V_i=\{(v_i)_{i\in I}:v_i\in V_i\},$$
$$\bigoplus_{i\in I}V_i=\Big\{(v_i)_{i\in I}\in\prod_{i\in I}V_i:v_i\neq 0\text{ for only finitely many }i\Big\},$$

with componentwise operations; the direct sum is a subspace of the direct product. For finite I the two coincide.

Example 4.5.24

Taking $V_n = k$ for every $n \in \mathbb{N}$, we get $\prod_{n\in\mathbb{N}} k = k^{\infty} \cong k[\![x]\!]$, while $\bigoplus_{n\in\mathbb{N}} k$ is the space of eventually zero sequences, isomorphic to $k[x]$ (Example 4.5.14).

4.6 Span, Linear Independence, and Bases

Span, linear independence and bases are the building blocks of linear algebra. They describe vector spaces in terms of generating sets and coordinates. We define them both for finite lists of vectors $(v_1, \dots, v_n)$, where order matters and repetitions are allowed (this is what coordinates and matrices need), and for arbitrary, possibly infinite, subsets $S \subseteq V$.

4.6.1 Linear Combinations and Span

Definition 4.6.1 (Linear combination)

A ***linear combination*** of vectors $v_1, \dots, v_n \in V$ is a vector of the form

$$a_1 v_1 + a_2 v_2 + \cdots + a_n v_n$$

with scalars $a_1, \dots, a_n \in k$. For $n = 0$ the empty sum is 0.

Definition 4.6.2 (Span)

The ***span*** of a subset $S \subseteq V$ is the set of all linear combinations of finitely many elements of S:

$$\operatorname{Span}(S) = \left\{ \sum_{i=1}^{n} a_i v_i : n \geqslant 0,\ a_i \in k,\ v_i \in S \right\}.$$

For a list $v_1, \dots, v_n$ we write $\operatorname{Span}(v_1, \dots, v_n) = \operatorname{Span}(\{v_1, \dots, v_n\}) = \{a_1 v_1 + \cdots + a_n v_n : a_i \in k\}$. We say S (or the list) ***spans*** V, or is a ***spanning set*** for V, if its span is V.

Even when S is infinite, only *finite* linear combinations are allowed; infinite sums have no meaning in a general vector space. Since the empty sum is 0, we have $0 \in \operatorname{Span}(S)$ always, and $\operatorname{Span}(\emptyset) = \{0\}$.

Proposition 4.6.3 (The span is the smallest subspace containing S)
For any subset $S \subseteq V$, $\mathrm{Span}(S)$ is a subspace of V that contains S, and it is contained in every subspace of V that contains S.

Proof. We have $0 \in \mathrm{Span}(S)$. A sum of two finite linear combinations of elements of S is again one, and so is a scalar multiple of one: $c\sum a_i v_i + c'\sum a'_i v'_i$ is a linear combination of the finitely many vectors v_i, v'_i. So $\mathrm{Span}(S)$ is a subspace. Each $v \in S$ equals $1 \cdot v \in \mathrm{Span}(S)$. Finally, a subspace $W \supseteq S$ is closed under addition and scalar multiplication, hence (by induction on n) contains every linear combination $\sum_{i=1}^{n} a_i v_i$ with $v_i \in S$. □

Equivalently, $\mathrm{Span}(S)$ is the intersection of all subspaces containing S, which is a subspace by Proposition 4.5.11.

4.6.2 Linear Independence

Definition 4.6.4 (Linear independence)

A list of vectors $v_1, \ldots, v_n \in V$ is ***linearly independent*** if the only way to write 0 as a linear combination of them is the trivial way:

$$a_1 v_1 + a_2 v_2 + \cdots + a_n v_n = 0 \implies a_1 = a_2 = \cdots = a_n = 0.$$

A subset $S \subseteq V$ is ***linearly independent*** if every list of finitely many *distinct* elements of S is linearly independent: for all $n \geqslant 1$, all distinct $v_1, \ldots, v_n \in S$ and all $a_1, \ldots, a_n \in k$,

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

A list or subset that is not linearly independent is ***linearly dependent***: some nontrivial finite linear combination of its elements is 0.

NOTE The word "distinct" matters. Without it, the relation $1 \cdot v + (-1) \cdot v = 0$ would make every nonempty set dependent. A list with a repeated vector, such as (v, v), is linearly dependent, while the set $\{v, v\} = \{v\}$ is linearly independent whenever $v \neq 0$. For a list $v_1, \ldots, v_n$ of distinct vectors, the list is independent exactly when the set $\{v_1, \ldots, v_n\}$ is. The empty list and the empty set are linearly independent.

NOTE Linear dependence means that some vector is a linear combination of the others. If $a_1v_1 + \cdots + a_nv_n = 0$ with, say, $a_1 \neq 0$, then

$$v_1 = -\frac{a_2}{a_1}v_2 - \cdots - \frac{a_n}{a_1}v_n,$$

so v_1 is redundant: it already lies in the span of the others.

Proposition 4.6.5 (Characterizations of linear independence)

The following are equivalent for vectors $v_1, \ldots, v_n \in V$:

(a) $v_1, \ldots, v_n$ are linearly independent;

(b) no v_j is a linear combination of the others;

(c) the linear map $\Phi : k^n \to V$, $\Phi(a_1, \ldots, a_n) = \sum a_iv_i$, is injective.

Proof. (a) $\Rightarrow$ (c): If $\Phi(a) = \Phi(b)$, then $\sum(a_i - b_i)v_i = 0$, so $a_i = b_i$ for all i by independence.

(c) $\Rightarrow$ (a): If $\sum a_iv_i = 0$, then $\Phi(a_1, \ldots, a_n) = 0 = \Phi(0, \ldots, 0)$, so all $a_i = 0$ by injectivity.

(a) $\Rightarrow$ (b): Suppose $v_j = \sum_{i \neq j} b_iv_i$. Then $\sum_{i \neq j} b_iv_i + (-1)v_j = 0$ is a nontrivial relation, since the coefficient of v_j is $-1 \neq 0$, contradicting independence.

(b) $\Rightarrow$ (a): Suppose $\sum a_iv_i = 0$ with some $a_j \neq 0$. Then $v_j = -a_j^{-1}\sum_{i \neq j} a_iv_i$, contradicting (b). $\square$

Lemma 4.6.6 (Adjoining a vector outside the span)

If $v_1, \ldots, v_n$ are linearly independent and $v \notin \operatorname{Span}(v_1, \ldots, v_n)$, then $v_1, \ldots, v_n, v$ are linearly independent.

Proof. Suppose $a_1v_1 + \cdots + a_nv_n + av = 0$. If $a \neq 0$, then $v = -a^{-1}(a_1v_1 + \cdots + a_nv_n) \in \operatorname{Span}(v_1, \ldots, v_n)$, a contradiction. So $a = 0$, and then $a_1 = \cdots = a_n = 0$ by independence of the v_i. $\square$

Proposition 4.6.7 (Characterization of spanning)

The following are equivalent for vectors $v_1, \ldots, v_n \in V$:

(a) $v_1, \ldots, v_n$ span V;

(b) every $v \in V$ can be written as $\sum a_iv_i$ for some $a_i \in k$;

(c) the linear map $\Phi : k^n \to V$, $\Phi(a_1, \ldots, a_n) = \sum a_iv_i$, is surjective.

Proof. Statement (b) is the definition of (a), and (c) says that every $v \in V$ is of the form $\Phi(a_1, \dots, a_n) = \sum a_i v_i$, which is (b). □

4.6.3 Bases

Definition 4.6.8 (Basis)

A ***basis*** of V is a linearly independent subset $\mathcal{B} \subseteq V$ that spans V. Similarly, a list $(v_1, \dots, v_n)$ is a basis of V if it is linearly independent and spans V; equivalently, the linear map $\Phi : k^n \to V$, $(a_1, \dots, a_n) \mapsto \sum a_i v_i$, is bijective.

When we speak of a finite basis we usually mean a list, so that its elements come in a definite order.

Proposition 4.6.9 (Unique representation)

A subset $\mathcal{B} \subseteq V$ is a basis if and only if every $v \in V$ can be written *uniquely* as a linear combination of distinct elements of $\mathcal{B}$; more precisely, there is a unique family of scalars $(a_b)_{b \in \mathcal{B}}$, only finitely many of them nonzero, with $v = \sum_{b \in \mathcal{B}} a_b b$. In particular, if $(v_1, \dots, v_n)$ is a basis of V, then every $v \in V$ can be written uniquely as

$$v = a_1 v_1 + a_2 v_2 + \cdots + a_n v_n$$

with $a_i \in k$. The scalars $(a_1, \dots, a_n)$ are the ***coordinates*** of v with respect to this basis.

Proof. Suppose $\mathcal{B}$ is a basis. Existence of an expression is the spanning property. If $v = \sum_b a_b b = \sum_b a'_b b$, then $\sum_b (a_b - a'_b) b = 0$, a finite linear combination of distinct elements of $\mathcal{B}$; by independence, $a_b = a'_b$ for all b.

Conversely, suppose every vector has a unique expression. Then $\mathcal{B}$ spans V. If $\sum_{i=1}^{n} a_i b_i = 0$ with distinct $b_i \in \mathcal{B}$, this is an expression of 0; comparing it with the expression in which all coefficients are 0, uniqueness gives $a_i = 0$ for all i. So $\mathcal{B}$ is independent. The statement for a list $(v_1, \dots, v_n)$ is proved in the same way. □

NOTE For an arbitrary subset $S \subseteq V$, let $k^{(S)}$ be the space of functions $a : S \to k$ with $a(s) \neq 0$ for only finitely many s. The map $k^{(S)} \to V$, $a \mapsto \sum_{s \in S} a(s)\, s$, is linear; S spans V if and only if it is surjective, S is linearly independent if and only if it is injective, and S is a basis if and only if it is bijective. For $S = \{v_1, \dots, v_n\}$ with distinct v_i, this is the map $\Phi : k^n \to V$ above.

Example 4.6.10

Standard basis of k^n: The vectors

$$e_1 = (1, 0, \ldots, 0), \quad e_2 = (0, 1, 0, \ldots, 0), \quad \ldots, \quad e_n = (0, \ldots, 0, 1)$$

form the ***standard basis*** of k^n: every vector $(a_1, \ldots, a_n) = a_1 e_1 + \cdots + a_n e_n$, and this is the only way to write it as a linear combination of the e_i.

Example 4.6.11

A non-standard basis of k^2: If $\mathrm{char}(k) \neq 2$, the vectors $(1, 1)$ and $(1, -1)$ also form a basis of k^2: every vector can be written as

$$(a, b) = \tfrac{a+b}{2}(1, 1) + \tfrac{a-b}{2}(1, -1),$$

and if $c(1, 1) + d(1, -1) = 0$, then $c + d = 0 = c - d$, so $2c = 0$ and $c = d = 0$. In characteristic 2 this fails: there $-1 = 1$, so $(1, -1) = (1, 1)$ and the list is linearly dependent.

Example 4.6.12

Bases of polynomial spaces: The polynomials $1, x, x^2, \ldots, x^n$ form a basis of $k[x]_{\leqslant n}$, and the infinite set $\{1, x, x^2, x^3, \ldots\}$ is a basis of $k[x]$. Indeed, every polynomial is a *finite* linear combination of monomials, and a linear combination $\sum_i a_i x^{n_i}$ of distinct monomials is the polynomial whose coefficients are the a_i, which is 0 only if every $a_i = 0$.

Example 4.6.13

In k^∞, let e_n be the sequence with 1 in position n and 0 elsewhere. The set $\{e_0, e_1, e_2, \ldots\}$ is linearly independent, but it spans only the eventually zero sequences: a finite linear combination of the e_n has only finitely many nonzero terms. So it is a basis of the subspace of eventually zero sequences, and not of k^∞.

4.7 Finite-Dimensional Vector Spaces and Dimension

The notion of *dimension* will appear constantly throughout linear algebra and beyond. Finite-dimensional vector spaces are determined up to isomorphism by a single integer, their dimension. We establish this through a sequence of lemmas about the interplay between spanning lists and linearly independent lists.

Definition 4.7.1 (Finite-dimensional)

A vector space V is ***finite-dimensional*** if it has a finite spanning list, i.e., vectors $v_1, \dots, v_n$ with $\operatorname{Span}(v_1, \dots, v_n) = V$. Otherwise, V is ***infinite-dimensional***.

4.7.1 Extracting Bases from Spanning Lists

Our first result shows that every finite spanning list contains a basis.

Lemma 4.7.2 (Spanning lists contain bases)

If $v_1, \dots, v_n$ span V, then some sublist of $v_1, \dots, v_n$ is a basis of V.

Proof. We construct a basis by removing redundant vectors one at a time.

If $v_1, \dots, v_n$ are linearly independent, they already form a basis. Otherwise, there is a nontrivial relation $\sum_{i=1}^{n} a_i v_i = 0$ with some $a_j \neq 0$. Then

$$v_j = -\frac{1}{a_j} \sum_{i \neq j} a_i v_i \in \operatorname{Span}(v_i : i \neq j).$$

So the list with v_j removed still spans V: in any linear combination of $v_1, \dots, v_n$ we can substitute this expression for v_j.

Remove v_j and repeat. Since the list is finite, the process terminates, and when it does, the remaining vectors are linearly independent and still span V, so they form a basis. (If $V = \{0\}$, we may end with the empty list, which is a basis of $\{0\}$.) □

Corollary 4.7.3 (Finite-dimensional spaces have bases)

Every finite-dimensional vector space has a finite basis.

4.7.2 Extending Linearly Independent Lists to Bases

The dual result shows that every linearly independent list can be enlarged to a basis.

Lemma 4.7.4 (Linearly independent lists extend to bases)

Let V be finite-dimensional and let $v_1, \dots, v_m \in V$ be linearly independent. Then $v_1, \dots, v_m$ can be extended to a basis of V. More precisely, if $w_1, \dots, w_r$ span V, then appending suitable vectors among $w_1, \dots, w_r$ to $v_1, \dots, v_m$ gives a basis of V.

Proof. Let $w_1, \dots, w_r$ span V; such a list exists because V is finite-dimensional. For $j = 0, 1, \dots, r$ let

$$W_j = \operatorname{Span}(v_1, \dots, v_m, w_1, \dots, w_j).$$

We show by induction on j that W_j has a basis consisting of $v_1, \dots, v_m$ followed by some of $w_1, \dots, w_j$. For $j = 0$, the list $v_1, \dots, v_m$ is independent and spans W_0, so it is a basis.

Suppose $\mathcal{B}_{j-1}$ is such a basis of W_{j-1}. If $w_j \in W_{j-1}$, then $W_j = W_{j-1}$ and we keep $\mathcal{B}_j = \mathcal{B}_{j-1}$. If $w_j \notin W_{j-1} = \operatorname{Span}(\mathcal{B}_{j-1})$, then $\mathcal{B}_j = (\mathcal{B}_{j-1}, w_j)$ is linearly independent by Lemma 4.6.6, and it spans W_j, since $\operatorname{Span}(\mathcal{B}_j)$ contains W_{j-1} and w_j. Either way $\mathcal{B}_j$ is a basis of W_j of the required form.

Finally, $W_r \supseteq \operatorname{Span}(w_1, \dots, w_r) = V$, so $\mathcal{B}_r$ is a basis of V containing $v_1, \dots, v_m$. □

4.7.3 The Exchange Lemma and Invariance of Dimension

The next lemma is the heart of the theory: it compares the lengths of independent lists and spanning lists. Its proof exchanges vectors of a spanning list, one at a time, for vectors of an independent list.

Lemma 4.7.5 (Exchange lemma)

If $u_1, \dots, u_m$ are linearly independent in V and $w_1, \dots, w_n$ span V, then $m \leqslant n$.

Proof. We show by induction on $j = 0, 1, \dots, m$ that $j \leqslant n$ and that, after renumbering the w's, the list

$$u_1, \dots, u_j, w_{j+1}, \dots, w_n$$

spans V. For $j = 0$ this is the hypothesis. Suppose it holds for some $j < m$. Since the list spans V, we can write

$$u_{j+1} = \sum_{i=1}^{j} b_i u_i + \sum_{l=j+1}^{n} c_l w_l.$$

If all the c_l were zero (in particular, if $j = n$ and no w's are left), then u_{j+1} would be a linear combination of $u_1, \dots, u_j$, contradicting the independence of the u's. So $j + 1 \leqslant n$, and after renumbering we may assume $c_{j+1} \neq 0$. Solving for w_{j+1},

$$w_{j+1} = c_{j+1}^{-1}\Big(u_{j+1} - \sum_{i=1}^{j} b_i u_i - \sum_{l=j+2}^{n} c_l w_l\Big) \in \operatorname{Span}(u_1, \dots, u_{j+1}, w_{j+2}, \dots, w_n).$$

So this span contains $u_1, \dots, u_j, w_{j+1}, \dots, w_n$, which span V; hence it is V. This completes the induction, and the case $j = m$ gives $m \leqslant n$. □

Theorem 4.7.6 (All bases have the same size)

If V has a basis with n elements, then every basis of V has exactly n elements.

Proof. Let $(v_1, \dots, v_n)$ be a basis and $\mathcal{B}'$ another basis of V. Any m distinct elements of $\mathcal{B}'$ are linearly independent, and $v_1, \dots, v_n$ span V, so $m \leqslant n$ by the exchange lemma. Hence $\mathcal{B}'$ is finite, with $m \leqslant n$ elements. Exchanging the roles of the two bases gives $n \leqslant m$. □

Definition 4.7.7 (Dimension)

The ***dimension*** of a finite-dimensional vector space V, denoted $\dim V$ or $\dim_k V$, is the number of elements of any basis of V. By the theorem this does not depend on the basis. The empty set is a basis of the zero space, so $\dim\{0\} = 0$. If V is infinite-dimensional we write $\dim V = \infty$.

Corollary 4.7.8 (Independent and spanning lists in dimension n)

Let $\dim V = n$.

(a) Every linearly independent list in V has at most n vectors, and every spanning list has at least n vectors.

(b) A linearly independent list of n vectors is a basis, and so is a spanning list of n vectors.

Proof. (a) Apply the exchange lemma with a basis of V as the spanning list, respectively as the independent list.

(b) By Lemma 4.7.4, an independent list of n vectors extends to a basis, which has n elements, so no vectors were added. By Lemma 4.7.2, a spanning list of n vectors contains a basis, which has n elements, so no vectors were removed. □

Corollary 4.7.9 (Classification of finite-dimensional spaces)

Every vector space V of dimension n is isomorphic to k^n. Two finite-dimensional vector spaces are isomorphic if and only if they have the same dimension.

Proof. Let $(v_1, \dots, v_n)$ be a basis of V. The map

$$\Phi : k^n \to V, \quad (a_1, \dots, a_n) \mapsto \sum_{i=1}^{n} a_i v_i$$

is linear, injective by Proposition 4.6.5 and surjective because the v_i span; so it is an isomorphism.

If $\varphi : V \to W$ is an isomorphism and $v_1, \dots, v_n$ is a basis of V, then $\varphi(v_1), \dots, \varphi(v_n)$ is a basis of W. It spans, since every $w \in W$ is $\varphi(v)$ for some $v = \sum a_i v_i$, so $w = \sum a_i \varphi(v_i)$. It is independent, since $\sum a_i \varphi(v_i) = 0$ means $\varphi(\sum a_i v_i) = 0 = \varphi(0)$, so $\sum a_i v_i = 0$ by injectivity and all $a_i = 0$. Hence isomorphic spaces have the same dimension. Conversely, two spaces of dimension n are both isomorphic to k^n, hence to each other. □

NOTE The isomorphism $\Phi : k^n \xrightarrow{\sim} V$ depends on the *choice of basis*, and different bases give different isomorphisms. Choosing a basis amounts to choosing coordinates: we represent $v = \sum a_i v_i$ by the ***column vector*** $\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \in k^n$ of its coordinates. Much of linear algebra consists in using coordinates when they help and avoiding them when they do not.

4.7.4 Subspaces of Finite-Dimensional Spaces

Proposition 4.7.10 (Subspaces of finite-dimensional spaces)

Let V be finite-dimensional and $U \subseteq V$ a subspace. Then U is finite-dimensional and $\dim U \leqslant \dim V$, with equality if and only if $U = V$.

Proof. Let $n = \dim V$. By the exchange lemma, every linearly independent list in U has at most n vectors, since it is also independent in V. Choose an independent list $u_1, \dots, u_m$ in U of the largest possible length; then $m \leqslant n$. It spans U: if some $u \in U$ were not in $\operatorname{Span}(u_1, \dots, u_m)$, then $u_1, \dots, u_m, u$ would be a longer independent list in U by Lemma 4.6.6. So $u_1, \dots, u_m$ is a basis of U, and $\dim U = m \leqslant n$.

If $U = V$, the dimensions agree. If $\dim U = n$, a basis of U is a linearly independent list of n vectors in V, hence a basis of V by Corollary 4.7.8; so $V = \operatorname{Span}(u_1, \dots, u_n) \subseteq U$, and $U = V$. □

Proposition 4.7.11 (Complements exist)

Let V be finite-dimensional and $U \subseteq V$ a subspace. Then there is a subspace $W \subseteq V$ with $V = U \oplus W$.

Proof. By Proposition 4.7.10, U has a basis $u_1, \dots, u_m$. Extend it to a basis $u_1, \dots, u_m, w_1, \dots, w_s$ of V (Lemma 4.7.4) and let $W = \operatorname{Span}(w_1, \dots, w_s)$. Every $v \in V$ is a linear combination of the u_i and w_j, so $V = U + W$. If $v \in U \cap W$, then $v = \sum a_i u_i = \sum b_j w_j$, so $\sum a_i u_i - \sum b_j w_j = 0$, and independence gives $a_i = b_j = 0$; thus $v = 0$. $\square$

A subspace W with $V = U \oplus W$ is called a ***complement*** of U in V. It is far from unique: in k^2, every line through the origin other than $U = \{(a, 0)\}$ is a complement of U.

4.7.5 Dimension of Sums and Direct Sums

Theorem 4.7.12 (Dimension formula for sums)

Let U and W be finite-dimensional subspaces of a vector space V. Then $U + W$ is finite-dimensional and

$$\dim(U + W) = \dim U + \dim W - \dim(U \cap W).$$

In particular, if the sum is direct, then $\dim(U \oplus W) = \dim U + \dim W$.

Proof. The subspace $U \cap W$ of U is finite-dimensional by Proposition 4.7.10. Let $v_1, \dots, v_r$ be a basis of $U \cap W$. By Lemma 4.7.4, extend it to a basis $v_1, \dots, v_r, u_1, \dots, u_m$ of U and to a basis $v_1, \dots, v_r, w_1, \dots, w_n$ of W. We claim that

$$v_1, \dots, v_r, u_1, \dots, u_m, w_1, \dots, w_n$$

is a basis of $U + W$.

Spanning: Any $x \in U + W$ is $x = u + w$ with $u \in U$ and $w \in W$. Writing u and w in the bases above expresses x as a linear combination of the listed vectors.

Independence: Suppose $\sum a_i v_i + \sum b_j u_j + \sum c_l w_l = 0$. Then

$$\sum c_l w_l = -\sum a_i v_i - \sum b_j u_j \in U.$$

Since also $\sum c_l w_l \in W$, it lies in $U \cap W$, so $\sum c_l w_l = \sum d_i v_i$ for some scalars d_i. Then $\sum d_i v_i - \sum c_l w_l = 0$, and since $v_1, \dots, v_r, w_1, \dots, w_n$ is a basis of W, all $c_l = 0$. Now $\sum a_i v_i + \sum b_j u_j = 0$, and since $v_1, \dots, v_r, u_1, \dots, u_m$ is a basis of U, all a_i and b_j vanish.

Thus $\dim(U + W) = r + m + n = (r + m) + (r + n) - r = \dim U + \dim W - \dim(U \cap W)$. □

Corollary 4.7.13 (Recognizing a direct sum decomposition)

Let V be finite-dimensional and $U, W \subseteq V$ subspaces. Then $V = U \oplus W$ if and only if $U \cap W = \{0\}$ and $\dim U + \dim W = \dim V$.

Proof. If $V = U \oplus W$, then $U \cap W = \{0\}$ and the formula gives $\dim V = \dim U + \dim W$. Conversely, under the two conditions the formula gives $\dim(U + W) = \dim V$, so $U + W = V$ by Proposition 4.7.10. □

Proposition 4.7.14 (Dimension of a direct sum)

(a) If $V = U_1 \oplus \cdots \oplus U_n$ and each U_i is finite-dimensional, then putting together bases of $U_1, \ldots, U_n$ gives a basis of V; so $\dim V = \dim U_1 + \cdots + \dim U_n$.

(b) If $V_1, \ldots, V_n$ are finite-dimensional, then $\dim(V_1 \times \cdots \times V_n) = \dim V_1 + \cdots + \dim V_n$.

Proof. (a) Let $u_{i,1}, \ldots, u_{i,d_i}$ be a basis of U_i. Every $v \in V$ is a sum $u_1 + \cdots + u_n$ with $u_i \in U_i$, and each u_i is a combination of $u_{i,1}, \ldots, u_{i,d_i}$; so the combined list spans V. If $\sum_{i,j} a_{ij} u_{i,j} = 0$, set $x_i = \sum_j a_{ij} u_{i,j} \in U_i$. Then $x_1 + \cdots + x_n = 0$, so every $x_i = 0$ by independence of the U_i, and then every $a_{ij} = 0$ by independence of the basis of U_i.

(b) The product $V_1 \times \cdots \times V_n$ is the internal direct sum of the subspaces $\{0\} \times \cdots \times V_i \times \cdots \times \{0\}$, each isomorphic to V_i and hence of dimension $\dim V_i$ (Corollary 4.7.9). Apply (a). Explicitly, if $v_{i,1}, \ldots, v_{i,d_i}$ is a basis of V_i, then the vectors $(0, \ldots, 0, v_{i,j}, 0, \ldots, 0)$, with $v_{i,j}$ in position i, form a basis of $V_1 \times \cdots \times V_n$. □

Example 4.7.15

Dimensions:

- $\dim k^n = n$: the standard basis has n elements.
- $\dim k[x]_{\leqslant d} = d + 1$: the monomials $1, x, \ldots, x^d$ form a basis.
- $k[x]$ is infinite-dimensional: any finite list of polynomials has bounded degree, so its span contains no polynomial of larger degree.

- $k[\![x]\!]$ and k^∞ are infinite-dimensional, since they contain the infinite-dimensional subspace $k[x]$, respectively the eventually zero sequences (Proposition 4.7.10).

Question 4.7.16

(a) Show that the monomials $1, x, x^2, \ldots$ are linearly independent in $k[\![x]\!]$ but do not span it.

(b) Suppose V has a countable spanning set. Show that every linearly independent subset of V is countable.

(c) Show that $k[\![x]\!]$ contains an uncountable linearly independent subset. Conclude that $k[\![x]\!]$ has no countable spanning set, and in particular no countable basis.

Solution. **(a)** A linear combination of finitely many distinct monomials is a polynomial, which is 0 only if all its coefficients are 0; so the monomials are independent. For the same reason their span is $k[x]$, and a power series such as $\sum_{i=0}^{\infty} x^i$, with infinitely many nonzero coefficients, is not in it.
(b) Let $s_0, s_1, s_2, \ldots$ span V (a finite spanning set can be padded with zeros) and put $V_n = \mathrm{Span}(s_0, \ldots, s_n)$. Every vector is a finite linear combination of the s_i, so $V = \bigcup_n V_n$. Let $I \subseteq V$ be linearly independent. Any distinct elements of $I \cap V_n$ form an independent list in a space spanned by $n+1$ vectors, so by the exchange lemma $I \cap V_n$ has at most $n+1$ elements. Hence $I = \bigcup_n (I \cap V_n)$ is a countable union of finite sets, and it is countable: list the elements of $I \cap V_0$, then the new elements of $I \cap V_1$, and so on.
(c) Let T be the set of finite strings of 0's and 1's, including the empty string. It is countable, since there are 2^l strings of each length l; fix a bijection $\beta : T \to \mathbb{N}$. For an infinite 0–1 sequence $t = (t_1, t_2, \ldots)$, let

$$A_t = \{\beta(t_1 t_2 \cdots t_l) : l \geqslant 0\} \subseteq \mathbb{N}$$

be the set of codes of the initial segments of t, and let $f_t = \sum_{n \in A_t} x^n \in k[\![x]\!]$. Each A_t is infinite, since initial segments of different lengths are different strings. If $t \neq t'$ and the two sequences first differ at position l_0, their common initial segments are those of length less than l_0, so $A_t \cap A_{t'}$ is finite. In particular $f_t \neq f_{t'}$.
The f_t are linearly independent. Suppose $\sum_{i=1}^{m} c_i f_{t_i} = 0$ with distinct $t_1, \ldots, t_m$. The set $A_{t_1} \setminus (A_{t_2} \cup \cdots \cup A_{t_m})$ is infinite, since it is obtained

from the infinite set A_{t_1} by removing the finite sets $A_{t_1} \cap A_{t_i}$; for n in it, the coefficient of x^n in $\sum c_i f_{t_i}$ is c_1. So $c_1 = 0$, and likewise every $c_i = 0$.
An infinite 0–1 sequence is a function $S \to \{0,1\}$ on the countably infinite set $S = \{1,2,3,\ldots\}$, and $|\{0,1\}^S| = |\mathcal{P}(S)| > |S|$ by Cantor's theorem (Theorem 1.7.7 and the discussion after it). So the set of infinite 0–1 sequences is uncountable, and since $t \mapsto f_t$ is injective, $\{f_t\}$ is an uncountable linearly independent subset of $k[\![x]\!]$. By (b), $k[\![x]\!]$ has no countable spanning set.
That $k[\![x]\!]$ has a basis at all follows from the theorem that every vector space has a basis, a consequence of Zorn's lemma that we do not prove. The proof of that theorem gives no way of writing a basis down.

CHAPTER 5

Linear Maps, Matrices, and Determinants

Having established the structure of vector spaces, we now study the maps between them. Linear maps were defined in Chapter 4; in this chapter we organize them into vector spaces of their own, represent them by matrices once bases are chosen, and see how the matrix changes when the bases change. We end with the determinant, the most important function of a square matrix, and with the determinant and trace of a linear operator. Throughout, k is a field.

5.1 Linear Maps

Recall (Definition 4.5.12) that a linear map $\varphi : V \to W$ between vector spaces over k is a function satisfying $\varphi(u+v) = \varphi(u) + \varphi(v)$ and $\varphi(\lambda v) = \lambda\varphi(v)$ for all $u, v \in V$ and $\lambda \in k$; it then preserves all linear combinations. Just as group homomorphisms preserve the group operation, linear maps preserve both vector space operations.

Example 5.1.1

- The zero map $v \mapsto 0$, the identity id_V, and the scaling $v \mapsto \lambda v$ for a fixed $\lambda \in k$ are linear.
- Differentiation $D : k[x] \to k[x]$, $D\big(\sum_i a_i x^i\big) = \sum_{i \geqslant 1} (i \cdot a_i) x^{i-1}$, is linear, since it acts coefficientwise. (This is the formal derivative; no limits are involved.)
- For $r \in k$, evaluation $k[x] \to k$, $f \mapsto f(r)$, is linear.
- If $v_1, \ldots, v_n$ is a basis of V, the coordinate isomorphism $k^n \to V$, $(a_1, \ldots, a_n) \mapsto \sum a_i v_i$, is linear.

- If $V = U \oplus W$, the ***projection*** onto U along W, sending $v = u + w$ ($u \in U$, $w \in W$) to u, is linear, because the decompositions of $v + v'$ and of λv are $(u + u') + (w + w')$ and $\lambda u + \lambda w$.

Proposition 5.1.2 (Linear maps are determined by a basis)

Let $v_1, \dots, v_n$ be a basis of V, and let $w_1, \dots, w_n \in W$ be arbitrary vectors. There is exactly one linear map $\varphi : V \to W$ with $\varphi(v_i) = w_i$ for all i, namely

$$\varphi\Big(\sum_{i=1}^{n} a_i v_i\Big) = \sum_{i=1}^{n} a_i w_i.$$

Proof. Uniqueness: a linear map with $\varphi(v_i) = w_i$ satisfies $\varphi\big(\sum a_i v_i\big) = \sum a_i \varphi(v_i) = \sum a_i w_i$.

Existence: every $v \in V$ has unique coordinates $a_1, \dots, a_n$ (Proposition 4.6.9), so the formula defines a function $\varphi : V \to W$, and $\varphi(v_i) = w_i$. If $u = \sum a_i v_i$ and $v = \sum b_i v_i$, then $u + v = \sum (a_i + b_i) v_i$ and $\lambda u = \sum (\lambda a_i) v_i$ are the coordinate expressions of $u + v$ and λu. Hence $\varphi(u + v) = \sum (a_i + b_i) w_i = \varphi(u) + \varphi(v)$ and $\varphi(\lambda u) = \lambda \varphi(u)$. □

NOTE This is the ***universal property*** of a basis: specifying a linear map out of V is the same as choosing where to send the basis vectors, and the choice is completely free. The same proof works for an arbitrary basis $\mathcal{B}$, finite or not: every function $\mathcal{B} \to W$ extends uniquely to a linear map $V \to W$.

Definition 5.1.3 (Space of linear maps)

The set of all linear maps from V to W is denoted

$$\mathrm{Hom}_k(V, W) \quad \text{or simply} \quad \mathrm{Hom}(V, W).$$

When $V = W$, we write $\mathrm{End}(V) = \mathrm{Hom}(V, V)$ and call its elements ***endomorphisms*** of V, or ***linear operators*** on V.

Proposition 5.1.4 ($\mathrm{Hom}(V, W)$ is a vector space)

The set $\mathrm{Hom}(V, W)$ is a vector space over k, with the operations

- **addition:** $(\varphi + \psi)(v) = \varphi(v) + \psi(v)$ for all $v \in V$;
- **scalar multiplication:** $(\lambda \varphi)(v) = \lambda \cdot \varphi(v)$ for all $v \in V$.

Proof. We must check that $\varphi + \psi$ and $\lambda\varphi$ are linear when φ and ψ are. Let $u, v \in V$ and $\mu \in k$. Then

$$(\varphi+\psi)(u+v) = \varphi(u)+\varphi(v)+\psi(u)+\psi(v) = (\varphi+\psi)(u)+(\varphi+\psi)(v),$$
$$(\varphi+\psi)(\mu v) = \mu\varphi(v)+\mu\psi(v) = \mu(\varphi+\psi)(v),$$

and similarly

$$(\lambda\varphi)(u+v) = \lambda(\varphi(u)+\varphi(v)) = (\lambda\varphi)(u)+(\lambda\varphi)(v),$$
$$(\lambda\varphi)(\mu v) = \lambda\mu\varphi(v) = \mu(\lambda\varphi)(v).$$

The vector space axioms for $\mathrm{Hom}(V, W)$ follow from those of W, because the operations are defined pointwise. The zero vector is the zero map $v \mapsto 0_W$, and the additive inverse of φ is $v \mapsto -\varphi(v)$. □

Proposition 5.1.5 (Composition)
Let $\varphi, \varphi' \in \mathrm{Hom}(U, V)$, $\psi, \psi' \in \mathrm{Hom}(V, W)$ and $\lambda \in k$. Then $\psi \circ \varphi$ is linear, and

$$\psi\circ(\varphi+\varphi') = \psi\circ\varphi+\psi\circ\varphi', \qquad (\psi+\psi')\circ\varphi = \psi\circ\varphi+\psi'\circ\varphi,$$
$$(\lambda\psi)\circ\varphi = \lambda(\psi\circ\varphi) = \psi\circ(\lambda\varphi).$$

Consequently $\mathrm{End}(V)$ is a ring, with addition as above, multiplication given by composition, and identity element id_V.

Proof. Composites of linear maps are linear by Proposition 4.5.13. For $u \in U$, linearity of ψ gives $\psi(\varphi(u)+\varphi'(u)) = \psi(\varphi(u))+\psi(\varphi'(u))$, which is the first identity; the second holds by the definition of $\psi+\psi'$; and the third follows from $\lambda\psi(\varphi(u)) = \psi(\lambda\varphi(u))$. For $\mathrm{End}(V)$: it is an abelian group under addition, composition of maps is associative with identity id_V, and the first two identities are the distributive laws. □

The ring $\mathrm{End}(V)$ is not commutative when $\dim V \geqslant 2$ (Example 5.2.11). Its units are the invertible linear operators on V, which form the group $\mathrm{GL}(V)$. We will see in Corollary 5.2.8 that $\dim \mathrm{Hom}(V, W) = (\dim V)(\dim W)$ when both spaces are finite-dimensional.

5.1.1 Dependence on the Scalar Field

The notions of basis, dimension and linearity depend on the choice of the field of scalars. The same set can be a vector space over different fields, with different dimensions.

Example 5.1.6

$\mathbb{C}$ as a vector space over different fields:

- Over $\mathbb{C}$: $\dim_{\mathbb{C}} \mathbb{C} = 1$, with basis $\{1\}$. Every complex number z is $z \cdot 1$.
- Over $\mathbb{R}$: $\dim_{\mathbb{R}} \mathbb{C} = 2$, with basis $\{1, i\}$. Every complex number $a + bi$ is the $\mathbb{R}$-linear combination $a \cdot 1 + b \cdot i$, and the expression is unique.

More generally, if k' is a subfield of k (for example $\mathbb{R} \subseteq \mathbb{C}$ or $\mathbb{Q} \subseteq \mathbb{R}$), then any k-vector space V is also a k'-vector space by ***restriction of scalars***: we only use scalar multiplication by elements of k'. In particular k itself is a k'-vector space. A k-linearly independent subset of V is also k'-linearly independent, since a k'-linear relation is in particular a k-linear relation; so $\dim_{k'} V \geqslant \dim_k V$.

Proposition 5.1.7 (Linearity over subfields)

Let V and W be k-vector spaces and $k' \subseteq k$ a subfield. Then:

(a) V and W are k'-vector spaces by restriction of scalars;

(b) every k-linear map $\varphi : V \to W$ is also k'-linear.

Thus $\mathrm{Hom}_k(V, W) \subseteq \mathrm{Hom}_{k'}(V, W)$.

Proof. The axioms for scalar multiplication by elements of k' are special cases of those for k, and $\varphi(\lambda v) = \lambda \varphi(v)$ for all $\lambda \in k$ holds in particular for $\lambda \in k'$. □

The inclusion $\mathrm{Hom}_k(V, W) \subseteq \mathrm{Hom}_{k'}(V, W)$ can be strict:

Example 5.1.8

Complex conjugation is $\mathbb{R}$-linear but not $\mathbb{C}$-linear. Consider

$$\mathbb{C} \to \mathbb{C}, \quad z = a + bi \mapsto \bar{z} = a - bi.$$

It is $\mathbb{R}$-linear: for $z_1, z_2 \in \mathbb{C}$ and $r \in \mathbb{R}$ we have $\overline{z_1 + z_2} = \bar{z}_1 + \bar{z}_2$ and $\overline{rz} = r\bar{z}$, since $\bar{r} = r$. It is not $\mathbb{C}$-linear: for $\lambda = i$ and $z = 1$,

$$\overline{i \cdot 1} = -i, \quad \text{but} \quad i \cdot \bar{1} = i.$$

So the natural maps between spaces depend on which scalars we allow. Complex conjugation is an $\mathbb{R}$-linear automorphism of $\mathbb{C}$ that is not $\mathbb{C}$-linear, because it does not commute with multiplication by i.

5.2 Matrices of Linear Maps

The correspondence between linear maps and matrices reduces questions about linear maps to computations with arrays of numbers.

5.2.1 Matrices and Matrix Multiplication

Definition 5.2.1 (Matrices)

An $m \times n$ ***matrix*** over k is an array $A = (a_{ij})$ of elements of k with m rows and n columns; a_{ij} is the entry in row i and column j. The set of $m \times n$ matrices is denoted $\mathcal{M}_{m\times n}(k)$, and $\mathcal{M}_n(k) = \mathcal{M}_{n\times n}(k)$. We identify k^n with the ***column vectors*** $\mathcal{M}_{n\times 1}(k)$, and write $A_j \in k^m$ for the j-th column of A. The ***transpose*** of A is the $n \times m$ matrix A^t with entries $(A^t)_{ij} = a_{ji}$.

With entrywise addition and scalar multiplication, $\mathcal{M}_{m\times n}(k)$ is a vector space; it is the space k^S of functions on $S = \{1,\dots,m\} \times \{1,\dots,n\}$. Let E_{ij} be the matrix with 1 in position (i,j) and 0 elsewhere. Every matrix can be written uniquely as $A = \sum_{i,j} a_{ij}E_{ij}$, so the mn matrices E_{ij} form a basis and

$$\dim \mathcal{M}_{m\times n}(k) = mn.$$

Definition 5.2.2 (Matrix product)

The product of $A \in \mathcal{M}_{m\times n}(k)$ and $B \in \mathcal{M}_{n\times p}(k)$ is the matrix $AB \in \mathcal{M}_{m\times p}(k)$ with entries

$$(AB)_{ij} = \sum_{l=1}^{n} a_{il}b_{lj}.$$

In particular, for a column vector $X \in k^n$, the product $AX \in k^m$ has entries $(AX)_i = \sum_{j=1}^{n} a_{ij}x_j$. The ***identity matrix*** is $I_n = (\delta_{ij})$, where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.

Some immediate consequences of the definition:

- $AX = x_1A_1 + \cdots + x_nA_n$ is the linear combination of the columns of A with coefficients x_j; in particular $Ae_j = A_j$.
- The j-th column of AB is AB_j, the matrix A applied to the j-th column of B.
- $I_mA = A = AI_n$ for $A \in \mathcal{M}_{m\times n}(k)$.
- $(AB)^t = B^tA^t$, since $((AB)^t)_{ij} = (AB)_{ji} = \sum_l a_{jl}b_{li} = \sum_l (B^t)_{il}(A^t)_{lj}$.
- The map $L_A : k^n \to k^m$, $X \mapsto AX$, is linear: $A(X+Y) = AX + AY$ and $A(\lambda X) = \lambda AX$.

5.2.2 The Matrix of a Linear Map

Let V and W be finite-dimensional, and fix bases $\mathcal{B} = (v_1, \dots, v_n)$ of V and $\mathcal{C} = (w_1, \dots, w_m)$ of W. Given a linear map $\varphi : V \to W$, we can express each image $\varphi(v_j)$ in the basis $\mathcal{C}$:

$$\varphi(v_j) = \sum_{i=1}^{m} a_{ij} w_i = a_{1j} w_1 + a_{2j} w_2 + \cdots + a_{mj} w_m.$$

Definition 5.2.3 (Matrix of a linear map)

The ***matrix of*** φ with respect to the bases $\mathcal{B}$ and $\mathcal{C}$ is the $m \times n$ matrix

$$\mathcal{M}(\varphi, \mathcal{B}, \mathcal{C}) = (a_{ij}) = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \in \mathcal{M}_{m \times n}(k),$$

whose j-th column is the coordinate vector of $\varphi(v_j)$ in the basis $\mathcal{C}$. For an operator $\varphi \in \mathrm{End}(V)$ we use the same basis on both sides and write $\mathcal{M}(\varphi, \mathcal{B}) = \mathcal{M}(\varphi, \mathcal{B}, \mathcal{B})$. When the bases are understood we simply write $\mathcal{M}(\varphi)$.

NOTE The key point: **the columns of the matrix are the images of the basis vectors, written in coordinates.** The j-th column tells you where the j-th basis vector goes.

5.2.3 Applying a Linear Map in Coordinates

Applying φ corresponds to multiplying by its matrix. Represent $x \in V$ by its coordinate vector: if $x = \sum_{j=1}^{n} x_j v_j$, let

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in k^n.$$

Similarly, let $y = \varphi(x) \in W$ have coordinates $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$, so that $y = \sum_{i=1}^{m} y_i w_i$.

Proposition 5.2.4 (Linear maps act by matrix multiplication)
With the notation above, and $A = \mathcal{M}(\varphi, \mathcal{B}, \mathcal{C})$, we have $Y = AX$, i.e.,

$$y_i = \sum_{j=1}^{n} a_{ij} x_j.$$

Proof. Compute:

$$\begin{aligned}\varphi(x) &= \varphi\left(\sum_{j=1}^{n} x_j v_j\right) = \sum_{j=1}^{n} x_j \varphi(v_j) = \sum_{j=1}^{n} x_j \left(\sum_{i=1}^{m} a_{ij} w_i\right) \\ &= \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij} x_j\right) w_i.\end{aligned}$$

Reading off the coefficient of w_i gives $y_i = \sum_{j=1}^{n} a_{ij} x_j$, the i-th entry of AX. □

In other words, if $\Phi_{\mathcal{B}} : k^n \to V$ and $\Phi_{\mathcal{C}} : k^m \to W$ are the coordinate isomorphisms of the two bases, the square

$$\begin{array}{ccc} V & \xrightarrow{\varphi} & W \\ {\scriptstyle \Phi_{\mathcal{B}}}\uparrow{\scriptstyle \cong} & & {\scriptstyle \cong}\uparrow{\scriptstyle \Phi_{\mathcal{C}}} \\ k^n & \xrightarrow[L_A]{} & k^m \end{array}$$

commutes: $\varphi \circ \Phi_{\mathcal{B}} = \Phi_{\mathcal{C}} \circ L_A$.

NOTE As a memory aid, the coordinate isomorphism $k^n \to V$ can be written symbolically as a row of vectors times a column of scalars:

$$(v_1\ v_2\ \cdots\ v_n) \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \sum_{j=1}^{n} x_j v_j.$$

(The v_j are vectors, so this is only notation.) Then φ acts as

$$\varphi\big((v_1\ \cdots\ v_n)X\big) = (w_1\ \cdots\ w_m)\,AX,$$

which makes the formula $Y = AX$ easy to remember.

Example 5.2.5

Let $D : k[x]_{\leqslant 3} \to k[x]_{\leqslant 3}$ be differentiation, and use the basis $(1, x, x^2, x^3)$. Since $D(1) = 0$, $D(x) = 1$, $D(x^2) = 2x$ and $D(x^3) = 3x^2$, the columns of the matrix are the coordinate vectors of these images:

$$\mathcal{M}(D) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

For instance, $f = 1 + x + x^2 + x^3$ has coordinates $(1,1,1,1)^t$, and $\mathcal{M}(D)(1,1,1,1)^t = (1,2,3,0)^t$ is the coordinate vector of $Df = 1 + 2x + 3x^2$.

Example 5.2.6

For $A \in \mathcal{M}_{m\times n}(k)$, the matrix of $L_A : k^n \to k^m$ in the standard bases is A itself, since $L_A(e_j) = Ae_j = A_j$.

5.2.4 The Isomorphism $\mathrm{Hom}(V, W) \cong \mathcal{M}_{m\times n}(k)$

Theorem 5.2.7 (Linear maps correspond to matrices)

Let $\dim V = n$ and $\dim W = m$, and fix bases $\mathcal{B}$ of V and $\mathcal{C}$ of W. The map

$$\mathrm{Hom}(V, W) \to \mathcal{M}_{m\times n}(k), \quad \varphi \mapsto \mathcal{M}(\varphi, \mathcal{B}, \mathcal{C})$$

sending a linear map to its matrix is an isomorphism of vector spaces.

Proof. Write $\mathcal{M}(\varphi)$ for $\mathcal{M}(\varphi, \mathcal{B}, \mathcal{C})$. We check that $\varphi \mapsto \mathcal{M}(\varphi)$ is linear, injective and surjective.

Linearity: Let $\varphi, \psi \in \mathrm{Hom}(V, W)$ and $\lambda \in k$. The j-th column of $\mathcal{M}(\varphi + \psi)$ is the coordinate vector of $(\varphi + \psi)(v_j) = \varphi(v_j) + \psi(v_j)$, which is the sum of the j-th columns of $\mathcal{M}(\varphi)$ and $\mathcal{M}(\psi)$. Hence $\mathcal{M}(\varphi + \psi) = \mathcal{M}(\varphi) + \mathcal{M}(\psi)$, and similarly $\mathcal{M}(\lambda\varphi) = \lambda\mathcal{M}(\varphi)$.

Injectivity: If $\mathcal{M}(\varphi) = 0$, then $\varphi(v_j) = 0$ for every basis vector v_j, so $\varphi = 0$ by Proposition 5.1.2.

Surjectivity: Given $A \in \mathcal{M}_{m\times n}(k)$, Proposition 5.1.2 gives a linear map φ with $\varphi(v_j) = \sum_i a_{ij} w_i$ for each j, and then $\mathcal{M}(\varphi) = A$. □

Corollary 5.2.8 (Dimension of $\mathrm{Hom}(V, W)$)

If $\dim V = n$ and $\dim W = m$, then

$$\dim \operatorname{Hom}(V, W) = \dim \mathcal{M}_{m\times n}(k) = mn.$$

Proof. Isomorphic spaces have the same dimension (Corollary 4.7.9), and the matrices E_{ij} form a basis of $\mathcal{M}_{m\times n}(k)$. □

NOTE Once we choose coordinates, linear maps *are* matrices, and the abstract theory of linear maps becomes concrete matrix computation. But the matrix depends on the bases: changing the bases changes the matrix, as we see in Section 5.3. One of our goals is to understand which properties of a linear map are intrinsic, that is, independent of the choice of basis.

5.2.5 Composition and Matrix Multiplication

Matrix multiplication is defined the way it is precisely so that it corresponds to composition of linear maps.

Proposition 5.2.9 (Matrix of a composite)

Let U, V, W be finite-dimensional with bases $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and let $\varphi : U \to V$ and $\psi : V \to W$ be linear. Then

$$\mathcal{M}(\psi \circ \varphi, \mathcal{A}, \mathcal{C}) = \mathcal{M}(\psi, \mathcal{B}, \mathcal{C})\, \mathcal{M}(\varphi, \mathcal{A}, \mathcal{B}).$$

Proof. Let $\mathcal{A} = (u_1, \dots, u_p)$, $\mathcal{B} = (v_1, \dots, v_n)$, $\mathcal{C} = (w_1, \dots, w_m)$, $A = \mathcal{M}(\varphi, \mathcal{A}, \mathcal{B})$ and $B = \mathcal{M}(\psi, \mathcal{B}, \mathcal{C})$. Then

$$\begin{aligned}(\psi \circ \varphi)(u_j) = \psi\Big(\sum_{l=1}^{n} a_{lj} v_l\Big) &= \sum_{l=1}^{n} a_{lj} \sum_{i=1}^{m} b_{il} w_i \\ &= \sum_{i=1}^{m} \Big(\sum_{l=1}^{n} b_{il} a_{lj}\Big) w_i = \sum_{i=1}^{m} (BA)_{ij} w_i.\end{aligned}$$

So the j-th column of $\mathcal{M}(\psi \circ \varphi)$ is the j-th column of BA. □

Corollary 5.2.10 (Algebra of matrices)

Matrix multiplication is associative, $(AB)C = A(BC)$, and distributes over addition, whenever the products are defined. In particular $\mathcal{M}_n(k)$ is a ring with identity I_n, and for any basis $\mathcal{B}$ of an n-dimensional space V, the map $\operatorname{End}(V) \to \mathcal{M}_n(k)$, $\varphi \mapsto \mathcal{M}(\varphi, \mathcal{B})$, is a ring isomorphism.

Proof. By Example 5.2.6, $A = \mathcal{M}(L_A)$ in the standard bases, so the proposition gives $\mathcal{M}(L_A \circ L_B) = AB$, that is, $L_A \circ L_B = L_{AB}$. Since composition of maps is associative, $L_{(AB)C} = L_A \circ L_B \circ L_C = L_{A(BC)}$, and taking matrices gives $(AB)C = A(BC)$. Distributivity follows in the same way from Proposition 5.1.5. The map $\varphi \mapsto \mathcal{M}(\varphi, \mathcal{B})$ is bijective and additive by Theorem 5.2.7, multiplicative by the proposition, and sends id_V to I_n. □

Example 5.2.11

Matrix multiplication is not commutative:

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Correspondingly, $\mathrm{End}(V)$ is not commutative when $\dim V \geqslant 2$: take the operators with these matrices on the span of two basis vectors, and zero on the remaining basis vectors.

Definition 5.2.12 (Invertible matrices)

A matrix $A \in \mathcal{M}_n(k)$ is ***invertible*** if there is $B \in \mathcal{M}_n(k)$ with $AB = BA = I_n$. Such a B is unique; it is the ***inverse*** A^{-1}. The invertible $n \times n$ matrices are the units of the ring $\mathcal{M}_n(k)$, and they form the ***general linear group*** $\mathrm{GL}_n(k)$.

Proposition 5.2.13 (Isomorphisms and invertible matrices)

Let $\varphi : V \to W$ be linear, where $\dim V = \dim W = n$, and let $\mathcal{B}, \mathcal{C}$ be bases of V, W. Then φ is an isomorphism if and only if $\mathcal{M}(\varphi, \mathcal{B}, \mathcal{C})$ is invertible, and in that case $\mathcal{M}(\varphi^{-1}, \mathcal{C}, \mathcal{B}) = \mathcal{M}(\varphi, \mathcal{B}, \mathcal{C})^{-1}$.

Proof. If φ is an isomorphism, φ^{-1} is linear (Proposition 4.5.13), and by Proposition 5.2.9

$$\mathcal{M}(\varphi^{-1}, \mathcal{C}, \mathcal{B})\,\mathcal{M}(\varphi, \mathcal{B}, \mathcal{C}) = \mathcal{M}(\mathrm{id}_V, \mathcal{B}, \mathcal{B}) = I_n,$$

and similarly in the other order. Conversely, suppose $A = \mathcal{M}(\varphi, \mathcal{B}, \mathcal{C})$ is invertible, and let $\psi : W \to V$ be the linear map with $\mathcal{M}(\psi, \mathcal{C}, \mathcal{B}) = A^{-1}$ (Theorem 5.2.7). Then $\mathcal{M}(\psi \circ \varphi, \mathcal{B}, \mathcal{B}) = A^{-1}A = I_n = \mathcal{M}(\mathrm{id}_V, \mathcal{B}, \mathcal{B})$, so $\psi \circ \varphi = \mathrm{id}_V$ by injectivity in Theorem 5.2.7. Similarly $\varphi \circ \psi = \mathrm{id}_W$. Hence φ is bijective. □

Proposition 5.2.14 (Invertibility and columns)

A matrix $A \in \mathcal{M}_n(k)$ is invertible if and only if its columns $A_1, \dots, A_n$ form a basis of k^n.

Proof. By Proposition 5.2.13 with the standard bases, A is invertible if and only if $L_A : k^n \to k^n$ is an isomorphism. If L_A is an isomorphism, it carries the standard basis to a basis (see the proof of Corollary 4.7.9), and $L_A(e_j) = A_j$. Conversely, if the columns form a basis, then $L_A(X) = \sum_j x_j A_j$ is the coordinate isomorphism of this basis, so L_A is an isomorphism. □

5.3 Change of Basis

What happens to the matrix of a linear map when we choose different bases? Let $\mathcal{B} = (v_1, \dots, v_n)$ and $\mathcal{B}' = (v'_1, \dots, v'_n)$ be two bases of V, the "old" and the "new" one. Write each new basis vector in terms of the old basis:

$$v'_j = \sum_{i=1}^{n} p_{ij} v_i.$$

Definition 5.3.1 (Change-of-basis matrix)

The $n \times n$ matrix $P = (p_{ij})$, whose j-th column is the coordinate vector of v'_j in the old basis $\mathcal{B}$, is the ***change-of-basis matrix*** from $\mathcal{B}'$ to $\mathcal{B}$. Comparing with the definition of the matrix of a linear map,

$$P = \mathcal{M}(\mathrm{id}_V, \mathcal{B}', \mathcal{B}),$$

the matrix of the identity map of V with the new basis used in the domain and the old basis in the target.

Symbolically,

$$(v'_1 \ \cdots \ v'_n) = (v_1 \ \cdots \ v_n)P, \qquad \text{so} \qquad (v'_1 \ \cdots \ v'_n)X' = (v_1 \ \cdots \ v_n)PX'.$$

In words:

Proposition 5.3.2 (Change of coordinates)

If a vector has coordinate vector X' in the new basis $\mathcal{B}'$, then its coordinate vector in the old basis $\mathcal{B}$ is $X = PX'$. The matrix P is invertible, with inverse $P^{-1} = \mathcal{M}(\mathrm{id}_V, \mathcal{B}, \mathcal{B}')$, and $X' = P^{-1}X$.

Proof. The first statement is Proposition 5.2.4 applied to id_V, with the basis $\mathcal{B}'$ in the domain and $\mathcal{B}$ in the target. By Proposition 5.2.9,

$$\mathcal{M}(\mathrm{id}_V, \mathcal{B}', \mathcal{B})\,\mathcal{M}(\mathrm{id}_V, \mathcal{B}, \mathcal{B}') = \mathcal{M}(\mathrm{id}_V, \mathcal{B}, \mathcal{B}) = I_n,$$

and likewise in the other order. □

Note the direction: P has the *new* basis vectors as its columns, but it converts *new* coordinates into *old* ones.

Theorem 5.3.3 (Change of basis)

Let $\varphi : V \to W$ be linear, let $\mathcal{B}, \mathcal{B}'$ be bases of V and $\mathcal{C}, \mathcal{C}'$ bases of W, and put

$$A = \mathcal{M}(\varphi, \mathcal{B}, \mathcal{C}), \qquad P = \mathcal{M}(\mathrm{id}_V, \mathcal{B}', \mathcal{B}), \qquad Q = \mathcal{M}(\mathrm{id}_W, \mathcal{C}, \mathcal{C}').$$

Then $\mathcal{M}(\varphi, \mathcal{B}', \mathcal{C}') = QAP$.

Proof. Write $\varphi = \mathrm{id}_W \circ \varphi \circ \mathrm{id}_V$ and apply Proposition 5.2.9, using the bases $\mathcal{B}' \to \mathcal{B} \to \mathcal{C} \to \mathcal{C}'$:

$$\mathcal{M}(\varphi, \mathcal{B}', \mathcal{C}') = \mathcal{M}(\mathrm{id}_W, \mathcal{C}, \mathcal{C}')\,\mathcal{M}(\varphi, \mathcal{B}, \mathcal{C})\,\mathcal{M}(\mathrm{id}_V, \mathcal{B}', \mathcal{B}) = QAP.$$

In the symbolic notation, with $(w_1 \ \cdots \ w_m) = (w'_1 \ \cdots \ w'_m)Q$:

$$\begin{aligned}\varphi\big((v'_1 \ \cdots \ v'_n)X'\big) &= \varphi\big((v_1 \ \cdots \ v_n)PX'\big)\\ &= (w_1 \ \cdots \ w_m)APX' = (w'_1 \ \cdots \ w'_m)QAPX'.\end{aligned}$$

□

Here Q is the inverse of the change-of-basis matrix from $\mathcal{C}'$ to $\mathcal{C}$. The most important case is an operator, with the same basis used on both sides.

Corollary 5.3.4 (Change of basis for operators)

Let $\varphi \in \mathrm{End}(V)$, let $\mathcal{B}$ and $\mathcal{B}'$ be bases of V, and let $P = \mathcal{M}(\mathrm{id}_V, \mathcal{B}', \mathcal{B})$. If $A = \mathcal{M}(\varphi, \mathcal{B})$ and $A' = \mathcal{M}(\varphi, \mathcal{B}')$, then

$$A' = P^{-1}AP.$$

Proof. Apply the theorem with $W = V$, $\mathcal{C} = \mathcal{B}$ and $\mathcal{C}' = \mathcal{B}'$. Then $Q = \mathcal{M}(\mathrm{id}_V, \mathcal{B}, \mathcal{B}') = P^{-1}$ by Proposition 5.3.2. □

So the matrices representing the *same* linear operator in different bases are related by ***conjugation*** by an invertible matrix. Still, the whole point of linear algebra is to work with linear maps in coordinate-free language as much as possible, and to choose bases only when they make a computation transparent.

5.3.1 Similar Matrices

Definition 5.3.5 (Similar matrices)

Two matrices $A, B \in \mathcal{M}_n(k)$ are ***similar***, written $A \sim B$, if there exists an invertible matrix $P \in GL_n(k)$ such that

$$B = P^{-1}AP.$$

Proposition 5.3.6 (Similarity is an equivalence relation)

Similarity is an equivalence relation on $\mathcal{M}_n(k)$.

Proof. It is reflexive: $A = I^{-1}AI$. It is symmetric: if $B = P^{-1}AP$, then $A = PBP^{-1} = (P^{-1})^{-1}B(P^{-1})$. It is transitive: if $B = P^{-1}AP$ and $C = Q^{-1}BQ$, then $C = Q^{-1}P^{-1}APQ = (PQ)^{-1}A(PQ)$. □

Proposition 5.3.7 (Similar matrices represent the same operator)

Two matrices $A, B \in \mathcal{M}_n(k)$ are similar if and only if there are an n-dimensional space V, an operator $\varphi \in \mathrm{End}(V)$ and bases $\mathcal{B}, \mathcal{B}'$ of V with $A = \mathcal{M}(\varphi, \mathcal{B})$ and $B = \mathcal{M}(\varphi, \mathcal{B}')$.

Proof. If A and B are matrices of one operator in two bases, they are similar by Corollary 5.3.4. Conversely, suppose $B = P^{-1}AP$. Take $V = k^n$ and $\varphi = L_A$, whose matrix in the standard basis $\mathcal{E}$ is A. The columns of P form a basis $\mathcal{B}'$ of k^n by Proposition 5.2.14, and $P = \mathcal{M}(\mathrm{id}, \mathcal{B}', \mathcal{E})$ because its columns are the vectors of $\mathcal{B}'$ written in the standard basis. By the corollary, $\mathcal{M}(L_A, \mathcal{B}') = P^{-1}AP = B$. □

So a similarity class of $n \times n$ matrices records exactly what one operator looks like in all possible bases. A property of matrices that is unchanged by similarity, a ***similarity invariant***, is therefore a property of the operator itself, shared by all the matrices that represent it. We now meet the first such invariant, the trace; the determinant follows in Section 5.4.

5.3.2 The Trace

Definition 5.3.8 (Trace of a matrix)

The ***trace*** of a square matrix $A = (a_{ij}) \in \mathcal{M}_n(k)$ is the sum of its diagonal entries:

$$\operatorname{Tr}(A) = \sum_{i=1}^{n} a_{ii} = a_{11} + a_{22} + \cdots + a_{nn}.$$

The trace is a linear map $\mathcal{M}_n(k) \to k$.

Lemma 5.3.9 (Cyclic property of the trace)

If $A \in \mathcal{M}_{m\times n}(k)$ and $B \in \mathcal{M}_{n\times m}(k)$, then $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$.

Proof. We have

$$\operatorname{Tr}(AB) = \sum_{i=1}^{m} (AB)_{ii} = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}b_{ji}, \qquad \operatorname{Tr}(BA) = \sum_{j=1}^{n} (BA)_{jj} = \sum_{j=1}^{n}\sum_{i=1}^{m} b_{ji}a_{ij}.$$

These are the same finite sum, taken in two different orders. □

Proposition 5.3.10 (Trace is a similarity invariant)

If $A \sim B$, then $\operatorname{Tr}(A) = \operatorname{Tr}(B)$.

Proof. If $B = P^{-1}AP$, then by the cyclic property applied to $P^{-1}A$ and P,

$$\operatorname{Tr}(B) = \operatorname{Tr}\left((P^{-1}A)P\right) = \operatorname{Tr}\left(P(P^{-1}A)\right) = \operatorname{Tr}(A).$$

□

Definition 5.3.11 (Trace of an operator)

Let V be finite-dimensional. The ***trace*** of an operator $\varphi \in \operatorname{End}(V)$ is

$$\operatorname{Tr}(\varphi) = \operatorname{Tr}\left(\mathcal{M}(\varphi, \mathcal{B})\right)$$

for any basis $\mathcal{B}$ of V. It does not depend on $\mathcal{B}$: by Corollary 5.3.4 the matrices of φ in two bases are similar, and similar matrices have the same trace.

Proposition 5.3.12 (Properties of the trace)

Let V be finite-dimensional, $\varphi, \psi \in \operatorname{End}(V)$ and $\lambda \in k$. Then $\operatorname{Tr}(\varphi + \psi) = \operatorname{Tr}(\varphi) + \operatorname{Tr}(\psi)$, $\operatorname{Tr}(\lambda\varphi) = \lambda \operatorname{Tr}(\varphi)$, $\operatorname{Tr}(\varphi \circ \psi) = \operatorname{Tr}(\psi \circ \varphi)$, and $\operatorname{Tr}(\mathrm{id}_V) = \dim V \cdot 1$.

Proof. Fix a basis and take matrices. The map $\varphi \mapsto \mathcal{M}(\varphi)$ is a ring isomorphism (Corollary 5.2.10), the trace of matrices is linear, $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$ by Lemma 5.3.9, and $\mathcal{M}(\mathrm{id}_V) = I_n$ has trace $n \cdot 1$. □

NOTE In Chapter 9 the trace of an operator is described without any choice of basis, as a contraction of tensors; that description makes its basis-independence evident.

5.4 Determinants

The determinant of a square matrix is a single scalar that detects whether the matrix is invertible, and that measures how the corresponding linear map scales volumes. We define it by an explicit formula, the Leibniz formula, and characterize it as the unique function of the columns of a matrix with three simple properties. Most of its other properties follow from this characterization. Throughout this section $n \geqslant 1$.

5.4.1 Alternating Multilinear Forms

Definition 5.4.1 (Multilinear and alternating forms)

Let V be a vector space over k and $d \geqslant 1$. A function $D : V^d = V \times \cdots \times V \to k$ is ***multilinear*** if it is linear in each argument when the others are held fixed:

$$D(v_1, \ldots, au + bw, \ldots, v_d) = a\, D(v_1, \ldots, u, \ldots, v_d) + b\, D(v_1, \ldots, w, \ldots, v_d)$$

for every position. It is ***alternating*** if $D(v_1, \ldots, v_d) = 0$ whenever $v_i = v_j$ for some $i \neq j$.

We use the sign homomorphism $\operatorname{sgn} : S_d \to \{1, -1\}$ of Chapter 3 (Definition 3.4.10). We need three facts about it from there: sgn is a group homomorphism, $\operatorname{sgn}(\tau) = -1$ for every transposition τ, and every permutation is a product of transpositions. Recall that permutations compose right to left: $\sigma\tau$ applies τ first.

Lemma 5.4.2 (Properties of alternating forms)
Let $D : V^d \to k$ be multilinear and alternating, and $v_1, \dots, v_d \in V$.

(a) Swapping two arguments changes the sign of D.
(b) For every $\sigma \in S_d$, $D(v_{\sigma(1)}, \dots, v_{\sigma(d)}) = \operatorname{sgn}(\sigma)\, D(v_1, \dots, v_d)$.
(c) If $v_1, \dots, v_d$ are linearly dependent, then $D(v_1, \dots, v_d) = 0$.
(d) Adding a multiple of one argument to a different argument does not change the value of D.

Proof. (a) Put $v + w$ in positions i and j, and expand by multilinearity in both positions (we display only these two):

$$0 = D(\dots, v+w, \dots, v+w, \dots) = D(\dots, v, \dots, v, \dots) + D(\dots, v, \dots, w, \dots) + D(\dots, w, \dots, v, \dots) + D(\dots, w, \dots, w, \dots).$$

The first and last terms vanish because D is alternating, so $D(\dots, w, \dots, v, \dots) = -D(\dots, v, \dots, w, \dots)$.

(b) Write σ as a product of r transpositions; we induct on r. For $r = 0$, $\sigma = \mathrm{id}$. If $r \geqslant 1$, write $\sigma = \tau\rho$ with $\tau = (i\ j)$ a transposition and ρ a product of $r-1$ transpositions. Put $w_l = v_{\tau(l)}$; then $(w_1, \dots, w_d)$ is $(v_1, \dots, v_d)$ with the entries in positions i and j swapped, and $v_{\sigma(l)} = v_{\tau(\rho(l))} = w_{\rho(l)}$. By induction applied to the vectors w_l, and by (a),

$$\begin{aligned} D(v_{\sigma(1)}, \dots, v_{\sigma(d)}) &= D(w_{\rho(1)}, \dots, w_{\rho(d)}) \\ &= \operatorname{sgn}(\rho)\, D(w_1, \dots, w_d) = -\operatorname{sgn}(\rho)\, D(v_1, \dots, v_d), \end{aligned}$$

and $-\operatorname{sgn}(\rho) = \operatorname{sgn}(\tau)\operatorname{sgn}(\rho) = \operatorname{sgn}(\sigma)$.

(c) Some v_j is a linear combination of the others, $v_j = \sum_{i \neq j} b_i v_i$ (Proposition 4.6.5). By linearity in the j-th argument,

$$D(v_1, \dots, v_d) = \sum_{i \neq j} b_i\, D(v_1, \dots, v_{j-1}, v_i, v_{j+1}, \dots, v_d),$$

where in the i-th term v_i occupies position j as well as position i. Each term vanishes because D is alternating.

(d) Replacing v_i by $v_i + c v_j$ ($j \neq i$) changes D by $c\, D(\dots, v_j, \dots, v_j, \dots)$, with v_j in positions i and j, which is 0. □

NOTE Property (a) is often taken as the definition of "alternating". If $\operatorname{char}(k) \neq 2$ the two versions agree: if swapping two arguments changes the sign, then $D(\dots, v, \dots, v, \dots) = -D(\dots, v, \dots, v, \dots)$, so

$2D(\dots,v,\dots,v,\dots) = 0$. In characteristic 2, where $-1 = 1$, property (a) says nothing, and the definition we have adopted is the right one.

5.4.2 The Determinant

We regard a matrix $A \in \mathcal{M}_n(k)$ as the list of its columns $A_1, \dots, A_n \in k^n$, and functions on $\mathcal{M}_n(k)$ as functions of n vectors in k^n. Accordingly, "multilinear and alternating in the columns" means multilinear and alternating as a function on $(k^n)^n$.

Definition 5.4.3 (Determinant)

The ***determinant*** of $A = (a_{ij}) \in \mathcal{M}_n(k)$ is

$$\det A = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma)\, a_{\sigma(1)1}\, a_{\sigma(2)2} \cdots a_{\sigma(n)n} = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{j=1}^{n} a_{\sigma(j)j}.$$

This is the ***Leibniz formula***. Each term picks one entry from every column, from pairwise different rows.

Theorem 5.4.4 (Characterization of the determinant)

(a) The determinant is multilinear and alternating in the columns, and $\det I_n = 1$.

(b) If $D : \mathcal{M}_n(k) \to k$ is multilinear and alternating in the columns, then $D(A) = D(I_n) \det A$ for all A.

In particular, det is the unique function $\mathcal{M}_n(k) \to k$ that is multilinear and alternating in the columns and takes the value 1 at I_n.

Proof. (b) Let D be multilinear and alternating in the columns, and write each column as $A_j = \sum_{i=1}^{n} a_{ij} e_i$. Expanding by linearity in each column in turn,

$$D(A) = D(A_1, \dots, A_n) = \sum_{i_1=1}^{n} \cdots \sum_{i_n=1}^{n} a_{i_1 1} a_{i_2 2} \cdots a_{i_n n}\, D(e_{i_1}, \dots, e_{i_n}).$$

The terms are indexed by the functions $f : \{1,\dots,n\} \to \{1,\dots,n\}$, $j \mapsto i_j$. If f is not injective, two arguments of $D(e_{i_1}, \dots, e_{i_n})$ coincide and the term vanishes. An injective map from $\{1,\dots,n\}$ to itself is a permutation σ, with $i_j = \sigma(j)$, and by Lemma 5.4.2(b), $D(e_{\sigma(1)}, \dots, e_{\sigma(n)}) = \operatorname{sgn}(\sigma) D(e_1, \dots, e_n) = \operatorname{sgn}(\sigma) D(I_n)$.

Hence

$$D(A) = D(I_n) \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{j=1}^{n} a_{\sigma(j)j} = D(I_n) \det A.$$

(a) *Multilinearity:* each product $\prod_j a_{\sigma(j)j}$ contains exactly one entry of the j-th column, so it is a linear function of A_j when the other columns are fixed; hence so is the sum.

Value at I_n*:* the entries of I_n are δ_{ij}, so $\prod_j \delta_{\sigma(j)j}$ is 1 if $\sigma = \mathrm{id}$ and 0 otherwise. Thus $\det I_n = \operatorname{sgn}(\mathrm{id}) = 1$.

Alternating: suppose columns $p \neq q$ of A are equal, and let $\tau = (p\ q)$. For $\sigma \in S_n$, compare the terms for σ and for $\sigma\tau$. For $j \notin \{p, q\}$ we have $\sigma\tau(j) = \sigma(j)$, while $a_{\sigma\tau(p)p} = a_{\sigma(q)p} = a_{\sigma(q)q}$ and $a_{\sigma\tau(q)q} = a_{\sigma(p)q} = a_{\sigma(p)p}$, because columns p and q are equal. So the two products are equal, while $\operatorname{sgn}(\sigma\tau) = -\operatorname{sgn}(\sigma)$. Since sgn is a homomorphism, $\sigma \mapsto \sigma\tau$ is a bijection from the even permutations to the odd ones, so

$$\det A = \sum_{\sigma \text{ even}} \left(\prod_j a_{\sigma(j)j} - \prod_j a_{\sigma\tau(j)j} \right) = 0.$$

The final statement combines (a) and (b). □

Example 5.4.5

- For $n = 1$: $\det(a) = a$.
- For $n = 2$: $S_2 = \{\mathrm{id}, (1\ 2)\}$, so

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc.$$

- For $n = 3$: the even permutations are id, $(1\ 2\ 3)$, $(1\ 3\ 2)$ and the odd ones are $(1\ 2)$, $(1\ 3)$, $(2\ 3)$. For instance $\sigma = (1\ 2\ 3)$ has $\sigma(1) = 2$, $\sigma(2) = 3$, $\sigma(3) = 1$ and contributes $+a_{21}a_{32}a_{13}$. Altogether,

$$\begin{aligned} \det A = {} & a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23} \\ & - a_{21}a_{12}a_{33} - a_{31}a_{22}a_{13} - a_{11}a_{32}a_{23}. \end{aligned}$$

In general the Leibniz formula has $n!$ terms, so it is a poor way to compute large determinants; the properties below give much better methods.

NOTE Over $\mathbb{R}$, $|\det A|$ is the area (for $n = 2$) or volume (for $n = 3$) of the parallelogram or parallelepiped spanned by the columns of A, and the sign records orientation. Multilinearity and the alternating property are

exactly how a signed volume should behave: scaling one edge scales the volume, and a degenerate parallelepiped, with two equal edges, has volume 0. Chapter 9 develops this point of view with volume forms.

5.4.3 Properties of the Determinant

Proposition 5.4.6 (Determinant of the transpose)
For every $A \in \mathcal{M}_n(k)$, $\det(A^t) = \det A$.

Proof. By the Leibniz formula applied to A^t, whose (i,j) entry is a_{ji},

$$\det(A^t) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{j=1}^{n} a_{j\sigma(j)}.$$

Reindex each product by $i = \sigma(j)$, so $j = \sigma^{-1}(i)$: then $\prod_j a_{j\sigma(j)} = \prod_i a_{\sigma^{-1}(i)i}$. Also $\operatorname{sgn}(\sigma^{-1}) = \operatorname{sgn}(\sigma)^{-1} = \operatorname{sgn}(\sigma)$, and $\sigma \mapsto \sigma^{-1}$ is a bijection of S_n. Hence

$$\det(A^t) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma^{-1}) \prod_{i=1}^{n} a_{\sigma^{-1}(i)i} = \sum_{\rho \in S_n} \operatorname{sgn}(\rho) \prod_{i=1}^{n} a_{\rho(i)i} = \det A.$$

□

Since the rows of A are the columns of A^t, everything we know about columns holds for rows as well: det is multilinear and alternating in the rows, swapping two rows changes the sign, adding a multiple of one row to another row does not change the determinant, and det is the unique function with these properties and $\det I_n = 1$. By multilinearity in all n columns, $\det(\lambda A) = \lambda^n \det A$ for $\lambda \in k$.

Theorem 5.4.7 (Multiplicativity)
For all $A, B \in \mathcal{M}_n(k)$,

$$\det(AB) = \det A \, \det B.$$

Proof. Fix A and define $D(B) = \det(AB)$. The j-th column of AB is AB_j, and $B_j \mapsto AB_j$ is linear; since det is multilinear in the columns, D is multilinear in the columns of B. If two columns of B are equal, so are the corresponding columns of AB, and $D(B) = 0$. By Theorem 5.4.4(b), $D(B) = D(I_n) \det B = \det(A) \det(B)$. □

Theorem 5.4.8 (Invertibility criterion)

A matrix $A \in \mathcal{M}_n(k)$ is invertible if and only if $\det A \neq 0$. In that case $\det(A^{-1}) = (\det A)^{-1}$.

Proof. If A is invertible, then $1 = \det I_n = \det(AA^{-1}) = \det(A)\det(A^{-1})$ by multiplicativity, so $\det A \neq 0$ and $\det(A^{-1}) = (\det A)^{-1}$. If A is not invertible, its columns do not form a basis of k^n (Proposition 5.2.14). Then they are linearly dependent, since n independent vectors in k^n would form a basis (Corollary 4.7.8). Hence $\det A = 0$ by Lemma 5.4.2(c). □

Corollary 5.4.9 (One-sided inverses)

If $A, B \in \mathcal{M}_n(k)$ and $AB = I_n$, then A and B are invertible and $B = A^{-1}$; in particular $BA = I_n$.

Proof. We have $\det(A)\det(B) = \det I_n = 1$, so $\det A \neq 0$ and A is invertible. Then $B = A^{-1}(AB) = A^{-1}$. □

Thus det restricts to a group homomorphism $GL_n(k) \to k^\times$. Its kernel is the ***special linear group*** $SL_n(k) = \{A \in GL_n(k) : \det A = 1\}$.

Proposition 5.4.10 (Triangular matrices)

If $A \in \mathcal{M}_n(k)$ is upper triangular ($a_{ij} = 0$ for $i > j$) or lower triangular ($a_{ij} = 0$ for $i < j$), then $\det A = a_{11}a_{22}\cdots a_{nn}$ is the product of the diagonal entries.

Proof. Let A be upper triangular. The term of the Leibniz formula for σ is nonzero only if $\sigma(j) \leqslant j$ for every j. Such a σ is the identity: $\sigma(1) \leqslant 1$ forces $\sigma(1) = 1$, and if $\sigma(i) = i$ for all $i < j$, then $\sigma(j) \leqslant j$ and $\sigma(j) \notin \{\sigma(1), \ldots, \sigma(j-1)\} = \{1, \ldots, j-1\}$ force $\sigma(j) = j$. So only the term $\sigma = \mathrm{id}$ survives, and it is $a_{11}\cdots a_{nn}$. A lower triangular matrix is the transpose of an upper triangular one with the same diagonal. □

Proposition 5.4.11 (Block triangular matrices)

Let $A \in \mathcal{M}_r(k)$, $D \in \mathcal{M}_s(k)$, and let B be $r \times s$ and C be $s \times r$. Then

$$\det \begin{pmatrix} A & B \\ 0 & D \end{pmatrix} = \det A \, \det D = \det \begin{pmatrix} A & 0 \\ C & D \end{pmatrix}.$$

Proof. Let M be the first matrix. Its first r columns are $\begin{pmatrix} A_j \\ 0 \end{pmatrix}$, which depend linearly on the columns A_j of A, and two equal columns of A give two equal columns of M. So, for fixed B and D, the function $A \mapsto \det M$ is multilinear and alternating in the columns of A, and Theorem 5.4.4(b) gives

$$\det M = \det A \cdot \det N, \qquad N = \begin{pmatrix} I_r & B \\ 0 & D \end{pmatrix}.$$

The $(r+l)$-th column of N is $\begin{pmatrix} B_l \\ D_l \end{pmatrix}$. Subtracting from it b_{il} times the i-th column $\begin{pmatrix} e_i \\ 0 \end{pmatrix}$, for $i = 1, \dots, r$, turns it into $\begin{pmatrix} 0 \\ D_l \end{pmatrix}$, and by Lemma 5.4.2(d) this does not change the determinant. So $\det N = \det \begin{pmatrix} I_r & 0 \\ 0 & D \end{pmatrix}$. This is multilinear and alternating in the columns of D, so by Theorem 5.4.4(b) again it equals $\det(I_n) \det D = \det D$. The second equality follows by taking transposes. □

By induction, the determinant of a block triangular matrix with square diagonal blocks $A_1, \dots, A_m$ is $\det A_1 \cdots \det A_m$.

Example 5.4.12

To compute a determinant efficiently, reduce to a triangular matrix with row operations, keeping track of their effect. For

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{pmatrix},$$

subtract 4 times row 1 from row 2 and 7 times row 1 from row 3, then 2 times the new row 2 from the new row 3:

$$\det A = \det \begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & -6 & -11 \end{pmatrix} = \det \begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & 0 & 1 \end{pmatrix} = 1 \cdot (-3) \cdot 1 = -3.$$

None of these operations changes the determinant. The Leibniz formula for $n = 3$ gives the same value: $50 + 96 + 84 - 80 - 105 - 48 = -3$.

5.4.4 Cofactor Expansion and the Adjugate

For $n \geqslant 2$ and $A \in \mathcal{M}_n(k)$, let $\widetilde{A}_{ij} \in \mathcal{M}_{n-1}(k)$ be the matrix obtained from A by deleting row i and column j.

Lemma 5.4.13 (Determinants with a standard basis column)

Let $n \geqslant 2$, and let N be the matrix obtained from A by replacing its j-th column with e_i. Then $\det N = (-1)^{i+j} \det \widetilde{A}_{ij}$.

Proof. Move the j-th column of N to the first position by swapping it successively with columns $j-1, \ldots, 1$; this is $j-1$ swaps of adjacent columns, and the other columns keep their order. Then move the i-th row to the first position by $i-1$ swaps of adjacent rows. Each swap multiplies the determinant by -1. The resulting matrix has first column e_1, and deleting its first row and column leaves $\widetilde{A}_{ij}$, since we deleted row i and column j of A and kept the others in order:

$$\begin{pmatrix} 1 & * \\ 0 & \widetilde{A}_{ij} \end{pmatrix}.$$

By Proposition 5.4.11 its determinant is $\det \widetilde{A}_{ij}$. Hence

$$\det N = (-1)^{(i-1)+(j-1)} \det \widetilde{A}_{ij} = (-1)^{i+j} \det \widetilde{A}_{ij}.$$

□

Theorem 5.4.14 (Cofactor expansion)

Let $n \geqslant 2$ and $A \in \mathcal{M}_n(k)$.

(a) For each column index j, $\det A = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det \widetilde{A}_{ij}$ (expansion along column j).

(b) For each row index i, $\det A = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det \widetilde{A}_{ij}$ (expansion along row i).

(c) For $l \neq j$ and for $l \neq i$, respectively,

$$\sum_{i=1}^{n} (-1)^{i+j} a_{il} \det \widetilde{A}_{ij} = 0, \qquad \sum_{j=1}^{n} (-1)^{i+j} a_{lj} \det \widetilde{A}_{ij} = 0.$$

Proof. (a) The j-th column of A is $\sum_i a_{ij} e_i$. By linearity of det in the j-th column and Lemma 5.4.13, $\det A = \sum_i a_{ij} (-1)^{i+j} \det \widetilde{A}_{ij}$.

(b) Apply (a) to A^t, expanding along its i-th column, and use $\widetilde{(A^t)}_{ji} = (\widetilde{A}_{ij})^t$ and Proposition 5.4.6.

(c) Let A' be the matrix obtained from A by replacing its j-th column with its l-th column. Then A' has two equal columns, so $\det A' = 0$, and $\widetilde{A'}_{ij} = \widetilde{A}_{ij}$ because column j is deleted. Expanding $\det A'$ along column j by (a) gives the first identity. The second follows in the same way from (b), replacing row i by row l. □

The scalar $(-1)^{i+j} \det \widetilde{A}_{ij}$ is the (i,j) ***cofactor*** of A. The signs $(-1)^{i+j}$ form a checkerboard pattern starting with $+$ in the top left corner.

Definition 5.4.15 (Adjugate)

The ***adjugate*** of $A \in \mathcal{M}_n(k)$, $n \geqslant 2$, is the matrix $\operatorname{adj}(A) \in \mathcal{M}_n(k)$ with entries

$$\operatorname{adj}(A)_{ji} = (-1)^{i+j} \det \widetilde{A}_{ij};$$

that is, $\operatorname{adj}(A)$ is the transpose of the matrix of cofactors. For $n = 1$ we set $\operatorname{adj}(a) = (1)$.

Theorem 5.4.16 (The adjugate identity)

For every $A \in \mathcal{M}_n(k)$,

$$\operatorname{adj}(A)\, A = A\, \operatorname{adj}(A) = \det(A)\, I_n.$$

Proof. For $n = 1$ this reads $1 \cdot a = a \cdot 1 = a$. Let $n \geqslant 2$. The (j,l) entry of $\operatorname{adj}(A)A$ is

$$\sum_{i=1}^{n} \operatorname{adj}(A)_{ji}\, a_{il} = \sum_{i=1}^{n} (-1)^{i+j} a_{il} \det \widetilde{A}_{ij},$$

which is $\det A$ if $l = j$ by Theorem 5.4.14(a), and 0 if $l \neq j$ by Theorem 5.4.14(c). Similarly, the (l,i) entry of $A \operatorname{adj}(A)$ is $\sum_j a_{lj} (-1)^{i+j} \det \widetilde{A}_{ij}$, which is $\det A$ if $l = i$ by (b) and 0 otherwise by (c). □

Corollary 5.4.17 (Formula for the inverse)

If $\det A \neq 0$, then $A^{-1} = (\det A)^{-1} \operatorname{adj}(A)$. In particular, for $ad - bc \neq 0$,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Proposition 5.4.18 (Cramer's rule)

Let $A \in \mathcal{M}_n(k)$ with $\det A \neq 0$, and $b \in k^n$. The unique solution $X \in k^n$ of $AX = b$ is given by

$$x_j = \frac{\det A^{(j)}}{\det A}, \qquad j = 1, \ldots, n,$$

where $A^{(j)}$ is the matrix obtained from A by replacing its j-th column with b.

Proof. The matrix A is invertible, so $X = A^{-1}b$ is the unique solution. Then $b = AX = \sum_i x_i A_i$. By linearity of det in the j-th column,

$$\det A^{(j)} = \sum_{i=1}^{n} x_i \det(A_1, \ldots, A_{j-1}, A_i, A_{j+1}, \ldots, A_n).$$

For $i \neq j$ the matrix in the i-th term has two equal columns, so only the term $i = j$ survives, and $\det A^{(j)} = x_j \det A$. □

5.4.5 The Determinant of an Operator

Proposition 5.4.19 (Determinant is a similarity invariant)
If $A \sim B$, then $\det A = \det B$.

Proof. If $B = P^{-1}AP$, then by multiplicativity

$$\det B = \det(P^{-1}) \det(A) \det(P) = (\det P)^{-1} \det(A) \det(P) = \det A.$$

□

Definition 5.4.20 (Determinant of an operator)

Let V be finite-dimensional. The ***determinant*** of $\varphi \in \mathrm{End}(V)$ is

$$\det \varphi = \det\big(\mathcal{M}(\varphi, \mathcal{B})\big)$$

for any basis $\mathcal{B}$ of V. It does not depend on $\mathcal{B}$, because the matrices of φ in two bases are similar (Corollary 5.3.4).

Proposition 5.4.21 (Properties of the determinant of an operator)
Let V be finite-dimensional of dimension n, and $\varphi, \psi \in \mathrm{End}(V)$, $\lambda \in k$. Then:

(a) $\det(\varphi \circ \psi) = \det \varphi \det \psi$ and $\det(\mathrm{id}_V) = 1$;
(b) φ is invertible if and only if $\det \varphi \neq 0$, and then $\det(\varphi^{-1}) = (\det \varphi)^{-1}$;
(c) $\det(\lambda\varphi) = \lambda^n \det \varphi$.

Proof. Fix a basis. By Corollary 5.2.10, $\mathcal{M}(\varphi \circ \psi) = \mathcal{M}(\varphi)\mathcal{M}(\psi)$ and $\mathcal{M}(\mathrm{id}_V) = I_n$, so (a) follows from Theorem 5.4.7. By Proposition 5.2.13, φ is invertible if and only if $\mathcal{M}(\varphi)$ is, so (b) follows from Theorem 5.4.8. Finally $\mathcal{M}(\lambda\varphi) = \lambda\mathcal{M}(\varphi)$ and $\det(\lambda A) = \lambda^n \det A$. □

NOTE In Chapter 9 we see the determinant of an operator without choosing a basis: φ acts on the top exterior power $\bigwedge^n V$ as multiplication by $\det \varphi$ (Theorem 9.11.3).

Example 5.4.22

Trace and determinant do not determine the similarity class of a matrix. Let

$$A = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}, \qquad B = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} = 2I_2$$

over $\mathbb{R}$. Both have trace 4 and determinant 4. But $P^{-1}BP = P^{-1}(2I_2)P = 2I_2 = B$ for every invertible P, so the similarity class of B is $\{B\}$, and $A \neq B$ is not similar to B.

NOTE Finer similarity invariants come from eigenvalues and eigenvectors, studied in Chapter 7. In the example, A and B have the same eigenvalue 2, but the vectors v with $Bv = 2v$ form all of $\mathbb{R}^2$, while those with $Av = 2v$ form only the line spanned by $(1,0)^t$. Over an algebraically closed field, the Jordan normal form of Chapter 7 (Theorem 7.8.4) classifies matrices up to similarity completely.

5.4.6 Permutation Matrices and the Sign

Definition 5.4.23 (Permutation matrix)

For $\sigma \in S_n$, let $T_\sigma : k^n \to k^n$ be the linear map with $T_\sigma(e_j) = e_{\sigma(j)}$ for all j (Proposition 5.1.2). Its matrix in the standard basis is the ***permutation matrix*** P_σ, whose j-th column is $e_{\sigma(j)}$; that is, $(P_\sigma)_{ij} = 1$ if $i = \sigma(j)$ and 0 otherwise.

Proposition 5.4.24 (Permutation matrices)

For $\sigma, \tau \in S_n$:

(a) $P_\sigma P_\tau = P_{\sigma\tau}$ and $P_{\mathrm{id}} = I_n$, so $\sigma \mapsto P_\sigma$ is an injective group homomorphism $S_n \to \mathrm{GL}_n(k)$;

(b) $\det P_\sigma = \operatorname{sgn}(\sigma)$.

Proof. (a) For each j, $T_\sigma(T_\tau(e_j)) = T_\sigma(e_{\tau(j)}) = e_{\sigma(\tau(j))} = T_{\sigma\tau}(e_j)$, so $T_\sigma \circ T_\tau = T_{\sigma\tau}$, and taking matrices gives $P_\sigma P_\tau = P_{\sigma\tau}$ (Corollary 5.2.10). Clearly $P_{\mathrm{id}} = I_n$, so

$P_\sigma P_{\sigma^{-1}} = I_n = P_{\sigma^{-1}} P_\sigma$ and each P_σ is invertible. If $P_\sigma = I_n$, then $e_{\sigma(j)} = e_j$ for all j, so $\sigma = \mathrm{id}$.

(b) The columns of P_σ are $e_{\sigma(1)}, \dots, e_{\sigma(n)}$, so by Lemma 5.4.2(b) applied to det, $\det P_\sigma = \operatorname{sgn}(\sigma) \det(e_1, \dots, e_n) = \operatorname{sgn}(\sigma)$. □

NOTE Over $\mathbb{R}$ one can see without computation that $\det P_\sigma = \pm 1$: the permutation σ has some finite order m, so $P_\sigma^m = P_{\sigma^m} = I_n$ by (a), hence $(\det P_\sigma)^m = 1$, and the only real roots of unity are ± 1. Part (b) says which sign occurs.

5.4.7 Determinants over Commutative Rings

The Leibniz formula makes sense for square matrices with entries in any commutative ring R (Definition 4.1.1), and so do the definitions of the matrix product, the transpose and the adjugate. The proofs of Lemma 5.4.2(a), (b) and (d), Theorem 5.4.4, Proposition 5.4.6, Theorem 5.4.7, Propositions 5.4.10 and 5.4.11, Lemma 5.4.13, Theorem 5.4.14 and Theorem 5.4.16 use only the ring axioms and commutativity: no division and no linear independence. (Associativity of matrix multiplication over R is a direct computation: the (i, j) entry of both $(AB)C$ and $A(BC)$ is $\sum_{l,m} a_{il} b_{lm} c_{mj}$.) So all these results hold verbatim for matrices over R, where "multilinear" means linear in each column with coefficients in R. In place of Theorem 5.4.8, a matrix A with entries in R has an inverse with entries in R if and only if $\det A$ is a unit of R: if $\det A \in R^\times$, then $(\det A)^{-1} \operatorname{adj}(A)$ is an inverse by the adjugate identity, and conversely $\det(A) \det(A^{-1}) = \det I_n = 1$.

Example 5.4.25

The case $R = k[x]$ is used in Chapter 7. For $A \in \mathcal{M}_n(k)$, the matrix $xI_n - A$ has entries in $k[x]$, so $\det(xI_n - A)$ is a polynomial in x, and $(xI_n - A) \operatorname{adj}(xI_n - A) = \det(xI_n - A)\, I_n$ holds in $\mathcal{M}_n(k[x])$. Moreover, if $f : R \to S$ is a ring homomorphism and $f(M)$ denotes the matrix obtained by applying f to every entry of M, then $f(\det M) = \det f(M)$, because $\det M$ is a sum of products of entries. Taking for f evaluation at $\lambda \in k$ shows that the value of the polynomial $\det(xI_n - A)$ at $x = \lambda$ is $\det(\lambda I_n - A)$. Taking for f complex conjugation shows $\det \overline{A} = \overline{\det A}$ for complex matrices, and hence $\det(A^*) = \overline{\det A}$ for the conjugate transpose $A^* = \overline{A}^{\,t}$.

Question 5.4.26

(a) Let $\varphi : \mathbb{R}^2 \to \mathbb{R}^2$ be the linear map with matrix $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ with respect to the standard basis. Find the matrix of φ with respect to the basis $\mathcal{B}' = ((1,1),(1,-1))$.

(b) Prove that $A \in \mathcal{M}_n(k)$ is similar to a diagonal matrix if and only if k^n has a basis $v_1, \ldots, v_n$ with $Av_i = \lambda_i v_i$ for some scalars $\lambda_i \in k$. (Such vectors are called eigenvectors of A; see Chapter 7.)

(c) We know $\operatorname{Tr}(A+B) = \operatorname{Tr}(A)+\operatorname{Tr}(B)$. Show that in general $\det(A+B) \neq \det A + \det B$.

(d) Compute the determinant and, using the adjugate, the inverse of

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 2 \end{pmatrix} \in \mathcal{M}_3(\mathbb{Q}).$$

Solution. **(a)** The change-of-basis matrix $P = \mathcal{M}(\mathrm{id}, \mathcal{B}', \mathcal{E})$ from $\mathcal{B}'$ to the standard basis $\mathcal{E}$ has the vectors of $\mathcal{B}'$ as its columns:

$$P = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \qquad P^{-1} = \frac{1}{-2}\begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix},$$

using the formula for the inverse of a 2×2 matrix. By Corollary 5.3.4, the matrix of φ with respect to $\mathcal{B}'$ is $A' = P^{-1}AP$. First,

$$AP = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 3 & -1 \\ 7 & -1 \end{pmatrix},$$

and then

$$A' = P^{-1}(AP) = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix}\begin{pmatrix} 3 & -1 \\ 7 & -1 \end{pmatrix} = \begin{pmatrix} 5 & -1 \\ -2 & 0 \end{pmatrix}.$$

As a check, $\operatorname{Tr}(A') = 5 = \operatorname{Tr}(A)$ and $\det(A') = 0 - 2 = -2 = 4 - 6 = \det(A)$, as similarity invariance requires.

(b) ($\Rightarrow$) Suppose $D = P^{-1}AP$ is diagonal, $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, and let $v_i = P_i$ be the i-th column of P. Then $AP = PD$, and comparing i-th columns gives $Av_i = \lambda_i v_i$. Since P is invertible, its columns form a basis of k^n (Proposition 5.2.14).

($\Leftarrow$) Suppose $v_1, \ldots, v_n$ is a basis with $Av_i = \lambda_i v_i$, and let P be the matrix with columns v_i. It is invertible by Proposition 5.2.14, and the i-th column of AP is $Av_i = \lambda_i v_i$, which is also the i-th column of PD for $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. So $AP = PD$ and $D = P^{-1}AP$.

(c) Take

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Then $\det A = \det B = 0$, but $A+B = I_2$, so $\det(A+B) = 1 \neq 0 = \det A + \det B$. So the determinant, although multiplicative and multilinear in the columns, is not additive.

(d) Expanding along the first row,

$$\det A = 2 \det \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix} - 1 \cdot \det \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix} + 0 = 2 \cdot 5 - 2 = 8.$$

The cofactors $(-1)^{i+j} \det \widetilde{A}_{ij}$ are

$$\begin{pmatrix} 5 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 5 \end{pmatrix};$$

for example, the $(2,1)$ cofactor is $-\det \begin{pmatrix} 1 & 0 \\ 1 & 2 \end{pmatrix} = -2$. This matrix is symmetric, so it equals its transpose $\operatorname{adj}(A)$, and

$$A^{-1} = \frac{1}{8} \begin{pmatrix} 5 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 5 \end{pmatrix}.$$

One checks directly that $A \operatorname{adj}(A) = 8I_3$; for instance, the first row of A times the first column of $\operatorname{adj}(A)$ is $2 \cdot 5 + 1 \cdot (-2) + 0 \cdot 1 = 8$, and times the second column is $2 \cdot (-2) + 1 \cdot 4 + 0 = 0$.

CHAPTER 6

Kernels, Images, and Duality

6.1 Kernel and Image

Every linear map carries two intrinsic subspaces that reveal its structure: the *kernel* (what gets annihilated) and the *image* (what gets hit). These are the linear-algebraic analogues of concepts we have seen before: the kernel of a group homomorphism and the image of a function. The relationship between kernel and image is governed by the ***Rank–Nullity Theorem***.

NOTE Kernel and image measure two different things. The kernel measures the *failure of injectivity*—how much information the map destroys. The image measures the *failure of surjectivity*—how much of the codomain the map reaches. The Rank–Nullity Theorem says these two failures are complementary: the more a linear map kills, the less it can produce, and vice versa.

Kernel and image do not determine the map. Every automorphism of V has kernel $\{0\}$ and image V; for instance id_V and $2 \cdot \mathrm{id}_V$ (when $2 \neq 0$ in k) have the same kernel and the same image. What is true is this: if U is a complement of $\ker(\varphi)$ in V (one exists when V is finite-dimensional, by Proposition 4.7.11), so that $V = \ker(\varphi) \oplus U$, then $\varphi(x + u) = \varphi(u)$ for $x \in \ker(\varphi)$ and $u \in U$, and φ restricts to an isomorphism $U \to \mathrm{Im}(\varphi)$ (its kernel is $U \cap \ker(\varphi) = \{0\}$). So φ is determined by its kernel together with this isomorphism, and knowing the kernel and the image leaves the isomorphism $U \to \mathrm{Im}(\varphi)$ undetermined.

6.1.1 The Kernel

Definition 6.1.1 (Kernel of a linear map)

Let $\varphi : V \to W$ be a linear map. The ***kernel*** of φ is

$$\ker(\varphi) = \{v \in V : \varphi(v) = 0_W\}.$$

This is the set of all vectors that φ maps to zero.

NOTE The term "kernel" comes from algebra (group homomorphisms), while "null space" is the synonym more common in applied linear algebra and matrix theory; Axler writes $\operatorname{null}\varphi$ for $\ker(\varphi)$. In Chapter 7 we will meet the *generalized kernel* of an operator $\varphi : V \to V$, the set of vectors killed by some power of φ. It contains $\ker(\varphi)$ and can be strictly larger.

Example 6.1.2 (Examples of kernels)

(a) The zero map $0 : V \to W$ has $\ker(0) = V$.

(b) The identity map $\mathrm{id}_V : V \to V$ has $\ker(\mathrm{id}_V) = \{0\}$.

(c) Let $\varphi : \mathbb{R}^3 \to \mathbb{R}$ be defined by $\varphi(x, y, z) = x + 2y + 3z$. Then $\ker(\varphi) = \{(x, y, z) \in \mathbb{R}^3 : x + 2y + 3z = 0\}$, a plane through the origin.

(d) Let $D : \mathbb{R}[x] \to \mathbb{R}[x]$ be differentiation, $Dp = p'$. Then $\ker(D)$ is the set of constant polynomials.

(e) The trace $\mathrm{Tr} : \mathcal{M}_{n\times n}(k) \to k$ of Chapter 5 is linear, and $\ker(\mathrm{Tr})$ is the space of matrices of trace zero.

Proposition 6.1.3 (The kernel is a subspace)

Let $\varphi : V \to W$ be a linear map. Then $\ker(\varphi)$ is a subspace of V.

Proof. We verify the subspace criteria:

(a) **Contains zero:** Since $\varphi(0_V) = 0_W$ (every linear map preserves zero), we have $0_V \in \ker(\varphi)$.

(b) **Closed under addition:** If $u, v \in \ker(\varphi)$, then $\varphi(u + v) = \varphi(u) + \varphi(v) = 0 + 0 = 0$, so $u + v \in \ker(\varphi)$.

(c) **Closed under scalar multiplication:** If $v \in \ker(\varphi)$ and $\lambda \in k$, then $\varphi(\lambda v) = \lambda\varphi(v) = \lambda \cdot 0 = 0$, so $\lambda v \in \ker(\varphi)$.

□

NOTE Compare this with group theory: the kernel of a group homomorphism $f : G \to H$ is a *normal* subgroup of G. For linear maps, every subspace is "normal" in the appropriate sense—vector spaces are abelian groups under addition, so there is no distinction between left and right cosets.

The kernel provides a clean characterization of injectivity:

Theorem 6.1.4 (Injectivity via the kernel)
Let $\varphi : V \to W$ be a linear map. Then

$$\varphi \text{ is injective} \iff \ker(\varphi) = \{0\}.$$

Proof. ($\Rightarrow$) Suppose φ is injective. Let $v \in \ker(\varphi)$. Then $\varphi(v) = 0 = \varphi(0)$. Since φ is injective, $v = 0$. Thus $\ker(\varphi) \subseteq \{0\}$, and since $0 \in \ker(\varphi)$ always, we have $\ker(\varphi) = \{0\}$.

($\Leftarrow$) Suppose $\ker(\varphi) = \{0\}$. Let $u, v \in V$ with $\varphi(u) = \varphi(v)$. Then

$$\varphi(u - v) = \varphi(u) - \varphi(v) = 0,$$

so $u - v \in \ker(\varphi) = \{0\}$. Thus $u - v = 0$, i.e., $u = v$. Hence φ is injective. □

NOTE The kernel does more than detect injectivity—it *measures the failure of injectivity* in a precise sense. Two vectors $u, v \in V$ have the same image under φ if and only if $u - v \in \ker(\varphi)$. Thus the fibers of φ (preimages of points of $\mathrm{Im}(\varphi)$) are exactly the cosets $v + \ker(\varphi)$. The map φ is injective precisely when these cosets are singletons, i.e., when $\ker(\varphi) = \{0\}$. This is why we can "mod out by the kernel" to obtain an injective map: the quotient space $V/\ker(\varphi)$, constructed in Section 7.1, collapses each fiber to a point.

6.1.2 The Image

Definition 6.1.5 (Image of a linear map)

Let $\varphi : V \to W$ be a linear map. The ***image*** (or ***range***) of φ is

$$\mathrm{Im}(\varphi) = \{\varphi(v) : v \in V\} = \{w \in W : w = \varphi(v) \text{ for some } v \in V\}.$$

This is the set of all vectors that φ "hits."

Example 6.1.6 (Examples of images)

(a) The zero map $0 : V \to W$ has $\mathrm{Im}(0) = \{0_W\}$.
(b) The identity map $\mathrm{id}_V : V \to V$ has $\mathrm{Im}(\mathrm{id}_V) = V$.
(c) Let $\pi : \mathbb{R}^3 \to \mathbb{R}^2$ be the projection $\pi(x, y, z) = (x, y)$. Then $\mathrm{Im}(\pi) = \mathbb{R}^2$.
(d) For $n \geqslant 1$, let $D : \mathbb{R}[x]_{\leqslant n} \to \mathbb{R}[x]_{\leqslant n-1}$ be differentiation. Then $\mathrm{Im}(D) = \mathbb{R}[x]_{\leqslant n-1}$: every polynomial of degree $\leqslant n-1$ is the derivative of a polynomial of degree $\leqslant n$.
(e) Let $A \in \mathcal{M}_{m\times n}(k)$ define $L_A : k^n \to k^m$, $L_A(x) = Ax$. Since $Ax = x_1A_1 + \cdots + x_nA_n$, where $A_1, \ldots, A_n$ are the columns of A, the image $\mathrm{Im}(L_A)$ is the ***column space*** of A, the span of its columns.

Proposition 6.1.7 (The image is a subspace)

Let $\varphi : V \to W$ be a linear map. Then $\mathrm{Im}(\varphi)$ is a subspace of W.

Proof. We verify the subspace criteria:

(a) **Contains zero:** $\varphi(0_V) = 0_W$, so $0_W \in \mathrm{Im}(\varphi)$.
(b) **Closed under addition:** If $w_1, w_2 \in \mathrm{Im}(\varphi)$, then $w_1 = \varphi(v_1)$ and $w_2 = \varphi(v_2)$ for some $v_1, v_2 \in V$. Thus $w_1 + w_2 = \varphi(v_1) + \varphi(v_2) = \varphi(v_1 + v_2) \in \mathrm{Im}(\varphi)$.
(c) **Closed under scalar multiplication:** If $w \in \mathrm{Im}(\varphi)$ and $\lambda \in k$, then $w = \varphi(v)$ for some $v \in V$, so $\lambda w = \lambda\varphi(v) = \varphi(\lambda v) \in \mathrm{Im}(\varphi)$.

□

Corollary 6.1.8

An injective linear map $\varphi : V \to W$ embeds V as a subspace of W: it is an isomorphism from V onto the subspace $\mathrm{Im}(\varphi)$ of W.

Proof. Regarded as a map $V \to \mathrm{Im}(\varphi)$, φ is linear and surjective, and it is injective by hypothesis. □

Definition 6.1.9 (Surjectivity)

A linear map $\varphi : V \to W$ is ***surjective*** (or ***onto***) if $\mathrm{Im}(\varphi) = W$, i.e., every element of W is hit by some element of V.

Proposition 6.1.10 (The image of a spanning set)

Let $\varphi : V \to W$ be a linear map and let S be a subset of V that spans V. Then

$$\mathrm{Im}(\varphi) = \mathrm{Span}(\varphi(S)), \qquad \varphi(S) = \{\varphi(s) : s \in S\}.$$

In particular, φ is surjective if and only if $\varphi(S)$ spans W, and if V is finite-dimensional, then so is $\mathrm{Im}(\varphi)$.

Proof. Since $\mathrm{Im}(\varphi)$ is a subspace containing $\varphi(S)$, it contains $\mathrm{Span}(\varphi(S))$. Conversely, every $v \in V$ is a linear combination $v = a_1 s_1 + \cdots + a_r s_r$ of elements of S, so $\varphi(v) = a_1\varphi(s_1) + \cdots + a_r\varphi(s_r) \in \mathrm{Span}(\varphi(S))$. If V is finite-dimensional, take S to be a finite basis; then $\mathrm{Im}(\varphi)$ is spanned by finitely many vectors. □

6.1.3 The Rank–Nullity Theorem

We now come to the central result connecting kernel and image. This theorem quantifies the fundamental trade-off in linear algebra: the more a map "kills," the less it can "produce."

Definition 6.1.11 (Rank and nullity)

Let $\varphi : V \to W$ be a linear map with V finite-dimensional. Then $\ker(\varphi)$ is finite-dimensional (it is a subspace of V; Proposition 4.7.10), and so is $\mathrm{Im}(\varphi)$ (Proposition 6.1.10). The ***rank*** of φ is

$$\mathrm{rank}(\varphi) = \dim(\mathrm{Im}(\varphi)),$$

and the ***nullity*** of φ is

$$\mathrm{null}(\varphi) = \dim(\ker(\varphi)).$$

For a matrix $A \in \mathcal{M}_{m\times n}(k)$, we define $\mathrm{rank}(A)$ and $\mathrm{null}(A)$ to be the rank and nullity of $L_A : k^n \to k^m$.

Theorem 6.1.12 (The Rank–Nullity Theorem)

Let $\varphi : V \to W$ be a linear map with V finite-dimensional. Then

$$\dim(V) = \dim(\ker(\varphi)) + \dim(\mathrm{Im}(\varphi)).$$

Equivalently: $\dim(V) = \mathrm{null}(\varphi) + \mathrm{rank}(\varphi)$. More precisely, if $u_1, \ldots, u_m$ is a basis of $\ker(\varphi)$ and $u_1, \ldots, u_m, v_1, \ldots, v_r$ is a basis of V, then $\varphi(v_1), \ldots, \varphi(v_r)$ is a basis of $\mathrm{Im}(\varphi)$.

Proof. The kernel is a subspace of the finite-dimensional space V, so it has a finite basis $\{u_1, \dots, u_m\}$ (Proposition 4.7.10), and every linearly independent list extends to a basis of V (Lemma 4.7.4). So bases as in the statement exist, and $\dim(V) = m + r$. It suffices to prove the last sentence.

Claim: $\{\varphi(v_1), \dots, \varphi(v_r)\}$ is a basis for $\mathrm{Im}(\varphi)$.

Spanning: Let $w \in \mathrm{Im}(\varphi)$, say $w = \varphi(x)$ for some $x \in V$. Write

$$x = a_1 u_1 + \cdots + a_m u_m + b_1 v_1 + \cdots + b_r v_r.$$

Applying φ:

$$w = \varphi(x) = a_1 \varphi(u_1) + \cdots + a_m \varphi(u_m) + b_1 \varphi(v_1) + \cdots + b_r \varphi(v_r).$$

Since each $u_i \in \ker(\varphi)$, we have $\varphi(u_i) = 0$, so

$$w = b_1 \varphi(v_1) + \cdots + b_r \varphi(v_r).$$

Thus $\{\varphi(v_1), \dots, \varphi(v_r)\}$ spans $\mathrm{Im}(\varphi)$.

Linear independence: Suppose

$$c_1 \varphi(v_1) + \cdots + c_r \varphi(v_r) = 0.$$

Then $\varphi(c_1 v_1 + \cdots + c_r v_r) = 0$, so $c_1 v_1 + \cdots + c_r v_r \in \ker(\varphi)$. Since $\{u_1, \dots, u_m\}$ is a basis for $\ker(\varphi)$, we can write

$$c_1 v_1 + \cdots + c_r v_r = d_1 u_1 + \cdots + d_m u_m$$

for some scalars d_j. Rearranging:

$$d_1 u_1 + \cdots + d_m u_m - c_1 v_1 - \cdots - c_r v_r = 0.$$

Since $\{u_1, \dots, u_m, v_1, \dots, v_r\}$ is linearly independent (it is a basis of V), all coefficients are zero. In particular, $c_1 = \cdots = c_r = 0$.

Thus $\{\varphi(v_1), \dots, \varphi(v_r)\}$ is a basis for $\mathrm{Im}(\varphi)$, so $\dim(\mathrm{Im}(\varphi)) = r$. We conclude:

$$\dim(V) = m + r = \dim(\ker(\varphi)) + \dim(\mathrm{Im}(\varphi)). \qquad \square$$

NOTE The Rank–Nullity Theorem is the linear-algebra counterpart of the First Isomorphism Theorem for groups (Theorem 3.3.3), which says $G/\ker(f) \cong \mathrm{Im}(f)$ for a homomorphism $f : G \to H$. In Section 7.1 we construct the quotient space $V/\ker(\varphi)$ and show $V/\ker(\varphi) \cong \mathrm{Im}(\varphi)$; Rank–Nullity is the statement that this isomorphism is compatible with counting dimensions, $\dim(V) - \dim(\ker(\varphi)) = \dim(\mathrm{Im}(\varphi))$. This is a way to remember the

theorem and is not a second proof of it: the formula $\dim(V/U) = \dim(V) - \dim(U)$ of Chapter 7 is itself deduced from Rank–Nullity.

6.1.4 A Normal Form for Linear Maps

The proof of the Rank–Nullity Theorem produces bases in which a linear map looks as simple as possible.

Theorem 6.1.13 (Normal form of a linear map)

Let $\varphi : V \to W$ be a linear map between finite-dimensional vector spaces, with $\dim(V) = n$, $\dim(W) = m$, and $r = \operatorname{rank}(\varphi)$. Then there exist a basis $\mathcal{B}$ of V and a basis $\mathcal{C}$ of W such that

$$M(\varphi, \mathcal{B}, \mathcal{C}) = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} \in \mathcal{M}_{m\times n}(k),$$

where I_r is the $r \times r$ identity matrix and the zeros are zero blocks of the appropriate sizes.

Proof. Choose a basis $u_1, \dots, u_{n-r}$ of $\ker(\varphi)$, and extend it by vectors $v_1, \dots, v_r$ to a basis of V. By Theorem 6.1.12, the vectors $w_1 = \varphi(v_1), \dots, w_r = \varphi(v_r)$ form a basis of $\operatorname{Im}(\varphi)$; extend them to a basis $w_1, \dots, w_m$ of W. Take $\mathcal{B} = (v_1, \dots, v_r, u_1, \dots, u_{n-r})$ and $\mathcal{C} = (w_1, \dots, w_m)$. For $j \leqslant r$, $\varphi(v_j) = w_j$, so the j-th column of the matrix is the j-th standard basis vector of k^m; the remaining columns are zero because $\varphi(u_i) = 0$. □

NOTE In this normal form the first r vectors of $\mathcal{C}$ span $\operatorname{Im}(\varphi)$ and the last $n-r$ vectors of $\mathcal{B}$ span $\ker(\varphi)$. The only information about φ left in the matrix is the number r: once we may choose bases of V and W independently, the rank is the only invariant of a linear map.

Corollary 6.1.14 (Matrices up to change of bases)

Let $A \in \mathcal{M}_{m\times n}(k)$ have rank r. Then there are invertible matrices $P \in \mathcal{M}_{n\times n}(k)$ and $Q \in \mathcal{M}_{m\times m}(k)$ with

$$Q^{-1}AP = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}.$$

Consequently, two matrices $A, A' \in \mathcal{M}_{m\times n}(k)$ satisfy $A' = Q^{-1}AP$ for some invertible P and Q if and only if $\operatorname{rank}(A) = \operatorname{rank}(A')$.

Proof. Apply Theorem 6.1.13 to $L_A : k^n \to k^m$, and let P and Q be the matrices whose columns are the vectors of $\mathcal{B}$ and of $\mathcal{C}$. By the change-of-basis formula (Theorem 5.3.3), the matrix of L_A in the bases $\mathcal{B}$ and $\mathcal{C}$ is $Q^{-1}AP$.

If $A' = Q^{-1}AP$, then $\operatorname{Im}(L_{AP}) = \operatorname{Im}(L_A)$ because L_P is surjective, and $L_{Q^{-1}}$ maps $\operatorname{Im}(L_{AP})$ isomorphically onto $\operatorname{Im}(L_{A'})$; hence $\operatorname{rank}(A') = \operatorname{rank}(A)$. Conversely, if both matrices have rank r, then $Q_1^{-1}AP_1 = Q_2^{-1}A'P_2$ for suitable invertible P_1, P_2, Q_1, Q_2, and $A' = (Q_1Q_2^{-1})^{-1}A(P_1P_2^{-1})$. □

6.1.5 Consequences of the Rank–Nullity Theorem

The Rank–Nullity Theorem has powerful implications for understanding linear maps.

Corollary 6.1.15 (Dimension constraints on injectivity)

Let $\varphi : V \to W$ be a linear map between finite-dimensional spaces with $\dim(V) = n$ and $\dim(W) = m$.

(a) If $n > m$, then φ cannot be injective.

(b) If $n < m$, then φ cannot be surjective.

Proof. (a) If φ were injective, then $\dim(\ker(\varphi)) = 0$, so by Rank–Nullity:

$$\dim(\operatorname{Im}(\varphi)) = \dim(V) = n > m = \dim(W).$$

But $\operatorname{Im}(\varphi) \subseteq W$, so $\dim(\operatorname{Im}(\varphi)) \leqslant \dim(W) = m$. Contradiction.

(b) If φ were surjective, then $\dim(\operatorname{Im}(\varphi)) = \dim(W) = m$, so by Rank–Nullity:

$$\dim(\ker(\varphi)) = \dim(V) - \dim(\operatorname{Im}(\varphi)) = n - m < 0.$$

But dimensions are non-negative. Contradiction. □

Corollary 6.1.16 (Equal dimensions: bijectivity criteria)

Let $\varphi : V \to W$ be a linear map with $\dim(V) = \dim(W) = n$ finite. The following are equivalent:

(a) φ is injective.

(b) φ is surjective.

(c) φ is bijective (an isomorphism).

(d) $\ker(\varphi) = \{0\}$.

(e) $\operatorname{Im}(\varphi) = W$.

(f) $\operatorname{rank}(\varphi) = n$.

Proof. By Rank–Nullity, $\operatorname{null}(\varphi) + \operatorname{rank}(\varphi) = n$. A subspace of W of dimension $n = \dim(W)$ is all of W, so $\operatorname{Im}(\varphi) = W$ if and only if $\operatorname{rank}(\varphi) = n$. Using Theorem 6.1.4:

- φ injective $\iff \ker(\varphi) = \{0\} \iff \operatorname{null}(\varphi) = 0 \iff \operatorname{rank}(\varphi) = n$.
- φ surjective $\iff \operatorname{Im}(\varphi) = W \iff \operatorname{rank}(\varphi) = n$.

Thus injectivity, surjectivity, and bijectivity are equivalent when dimensions match. □

NOTE For linear maps between equal-dimensional spaces, "one-to-one" and "onto" are the same condition. This fails for general functions and even for linear maps between infinite-dimensional spaces. Consider the ***right shift*** $R : k^\infty \to k^\infty$ defined by $R(x_1, x_2, x_3, \ldots) = (0, x_1, x_2, \ldots)$. This is injective (its kernel is trivial) but not surjective (the sequence $(1, 0, 0, \ldots)$ is not in the image). Conversely, the ***left shift*** $L(x_1, x_2, x_3, \ldots) = (x_2, x_3, \ldots)$ is surjective but not injective. In infinite dimensions there is "room" for one without the other. The finite-dimensional equivalence is the linear analogue of the pigeonhole principle.

Corollary 6.1.17 (Existence of left and right inverses)

Let $\varphi : V \to W$ be a linear map between finite-dimensional spaces.

(a) φ has a left inverse (a linear $\psi : W \to V$ with $\psi \circ \varphi = \mathrm{id}_V$) if and only if φ is injective.

(b) φ has a right inverse (a linear $\psi : W \to V$ with $\varphi \circ \psi = \mathrm{id}_W$) if and only if φ is surjective.

Proof. (a) If $\psi \circ \varphi = \mathrm{id}_V$ and $\varphi(v) = 0$, then $v = \psi(\varphi(v)) = 0$, so φ is injective. Conversely, let φ be injective and let $v_1, \ldots, v_n$ be a basis of V. The vectors $\varphi(v_1), \ldots, \varphi(v_n)$ are linearly independent: if $\sum c_i \varphi(v_i) = 0$, then $\varphi(\sum c_i v_i) = 0$, so $\sum c_i v_i = 0$ and all $c_i = 0$. Extend them to a basis $\varphi(v_1), \ldots, \varphi(v_n), w_1, \ldots, w_s$ of W, and define ψ on this basis by $\psi(\varphi(v_i)) = v_i$ and $\psi(w_j) = 0$ (a linear map may be prescribed arbitrarily on a basis, Proposition 5.1.2). Then $\psi \circ \varphi$ agrees with id_V on a basis, so $\psi \circ \varphi = \mathrm{id}_V$.

(b) If $\varphi \circ \psi = \mathrm{id}_W$, then every $w \in W$ equals $\varphi(\psi(w))$, so φ is surjective. Conversely, let φ be surjective and let $w_1, \ldots, w_m$ be a basis of W. Choose $v_j \in V$ with $\varphi(v_j) = w_j$, and let $\psi : W \to V$ be the linear map with $\psi(w_j) = v_j$. Then $\varphi \circ \psi$ agrees with id_W on a basis, so $\varphi \circ \psi = \mathrm{id}_W$. □

6.1.6 Matrix Interpretation

For a matrix $A \in \mathcal{M}_{m\times n}(k)$, the associated linear map is $L_A : k^n \to k^m$ given by $L_A(x) = Ax$. The Rank–Nullity Theorem becomes:

Proposition 6.1.18 (Rank–Nullity for matrices)

For $A \in \mathcal{M}_{m\times n}(k)$:

(a) $\ker(L_A) = \{x \in k^n : Ax = 0\}$ is the *null space* of A.

(b) $\mathrm{Im}(L_A) = \{Ax : x \in k^n\}$ is the *column space* of A (the span of the columns).

(c) $\mathrm{null}(A) + \mathrm{rank}(A) = n$ (the number of columns).

NOTE Since the columns of A span the column space and every spanning list contains a basis (Lemma 4.7.2), the rank of A is the largest number of linearly independent columns of A. It is also the largest number of linearly independent rows; this is Corollary 6.2.17 below, "row rank equals column rank." The null space has dimension $n - \mathrm{rank}(A)$.
Looking ahead to Chapter 7: for a linear operator $\varphi : V \to V$ and a scalar λ, the *eigenspace* $\ker(\varphi - \lambda \mathrm{id}_V)$ is the kernel of a shifted operator. Its dimension (the *geometric multiplicity* of λ) measures "how degenerate" $\varphi - \lambda\mathrm{id}_V$ is, and the gap between geometric and algebraic multiplicities is what makes the Jordan normal form necessary.

Example 6.1.19 (Computing kernel and image)

Let $A = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 1 & 3 & 5 \end{pmatrix}$ over $\mathbb{R}$.

Finding $\ker(A)$**:** Solve $Ax = 0$. Row operations do not change the solutions of $Ax = 0$, so we row reduce:

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 1 & 3 & 5 \end{pmatrix} \xrightarrow{R_3 - R_1} \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 1 & 2 \end{pmatrix} \xrightarrow{R_3 - R_2} \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

Back-substituting: $x_2 = -2x_3$, $x_1 = -2x_2 - 3x_3 = 4x_3 - 3x_3 = x_3$. So

$$\ker(A) = \mathrm{Span}\{(1, -2, 1)^t\}, \quad \mathrm{null}(A) = 1.$$

Finding $\mathrm{Im}(A)$**:** The echelon form has its pivots in columns 1 and 2. Since row operations preserve the linear relations among the columns (a vector

x satisfies $Ax = 0$ if and only if it satisfies $A'x = 0$ for the reduced matrix A'), columns 1 and 2 of A are linearly independent, and column 3 is a combination of them: $(3,2,5)^t = -(1,0,1)^t + 2\,(2,1,3)^t$. Hence

$$\operatorname{Im}(A) = \operatorname{Span}\left\{ \begin{pmatrix}1\\0\\1\end{pmatrix}, \begin{pmatrix}2\\1\\3\end{pmatrix} \right\}, \quad \operatorname{rank}(A) = 2.$$

Verification: $\operatorname{null}(A) + \operatorname{rank}(A) = 1 + 2 = 3 = n$. ✓

Question 6.1.20

(a) Let $D : \mathbb{R}[x]_{\leqslant 3} \to \mathbb{R}[x]_{\leqslant 3}$ be the differentiation operator. Find $\ker(D)$, $\operatorname{Im}(D)$, $\operatorname{null}(D)$, and $\operatorname{rank}(D)$. Verify the Rank–Nullity Theorem.

(b) Let $T : \mathbb{R}^4 \to \mathbb{R}^3$ be defined by $T(x_1, x_2, x_3, x_4) = (x_1 - x_2, x_2 - x_3, x_3 - x_4)$. Determine whether T is injective, surjective, or neither.

(c) Prove that if $\varphi : V \to V$ is a linear operator with $\varphi^2 = \varphi$, then $V = \ker(\varphi) \oplus \operatorname{Im}(\varphi)$.

(d) Prove that a linear map $\varphi : V \to W$ is injective if and only if for every linearly independent list in V, its image under φ is linearly independent in W.

Solution. **(a)** The space $\mathbb{R}[x]_{\leqslant 3}$ has basis $\{1, x, x^2, x^3\}$, so $\dim(\mathbb{R}[x]_{\leqslant 3}) = 4$.
Kernel: $Dp = 0$ if and only if p is a constant polynomial. So $\ker(D) = \operatorname{Span}\{1\}$, the constant polynomials. Thus $\operatorname{null}(D) = 1$.
Image: Differentiating: $D(1) = 0$, $D(x) = 1$, $D(x^2) = 2x$, $D(x^3) = 3x^2$. By Proposition 6.1.10, these span the image, so $\operatorname{Im}(D) = \operatorname{Span}\{1, x, x^2\} = \mathbb{R}[x]_{\leqslant 2}$ and $\operatorname{rank}(D) = 3$.
Verification: $\operatorname{null}(D) + \operatorname{rank}(D) = 1 + 3 = 4 = \dim(\mathbb{R}[x]_{\leqslant 3})$. ✓
(b) Injectivity: Find $\ker(T)$. Solve $T(x_1, x_2, x_3, x_4) = (0, 0, 0)$:

$$x_1 - x_2 = 0, \quad x_2 - x_3 = 0, \quad x_3 - x_4 = 0.$$

This gives $x_1 = x_2 = x_3 = x_4$. So $\ker(T) = \{(t, t, t, t) : t \in \mathbb{R}\} = \operatorname{Span}\{(1, 1, 1, 1)\}$.
Since $\ker(T) \neq \{0\}$, T is **not injective**.
Surjectivity: We have $\operatorname{null}(T) = 1$, so $\operatorname{rank}(T) = 4 - 1 = 3 = \dim(\mathbb{R}^3)$. Thus $\operatorname{Im}(T) = \mathbb{R}^3$, and T is **surjective**.
(c) We show $V = \ker(\varphi) + \operatorname{Im}(\varphi)$ and $\ker(\varphi) \cap \operatorname{Im}(\varphi) = \{0\}$.

Sum equals V: Let $v \in V$. Write $v = (v - \varphi(v)) + \varphi(v)$. Note that $\varphi(v) \in \mathrm{Im}(\varphi)$. Also:

$$\varphi(v - \varphi(v)) = \varphi(v) - \varphi^2(v) = \varphi(v) - \varphi(v) = 0,$$

so $v - \varphi(v) \in \ker(\varphi)$. Thus $v \in \ker(\varphi) + \mathrm{Im}(\varphi)$.
Intersection is trivial: Let $w \in \ker(\varphi) \cap \mathrm{Im}(\varphi)$. Then $\varphi(w) = 0$ and $w = \varphi(u)$ for some $u \in V$. So:

$$w = \varphi(u) = \varphi^2(u) = \varphi(\varphi(u)) = \varphi(w) = 0.$$

Thus $V = \ker(\varphi) \oplus \mathrm{Im}(\varphi)$.
(d) ($\Rightarrow$) Suppose φ is injective and $v_1, \ldots, v_m$ is linearly independent. Suppose $\sum c_i \varphi(v_i) = 0$. Then $\varphi(\sum c_i v_i) = 0$, so $\sum c_i v_i \in \ker(\varphi) = \{0\}$. Thus $\sum c_i v_i = 0$, and by linear independence of the v_i, all $c_i = 0$. Hence $\varphi(v_1), \ldots, \varphi(v_m)$ is linearly independent.
($\Leftarrow$) Suppose φ maps every linearly independent list to a linearly independent list. If $v \neq 0$, then the list (v) is linearly independent, so $(\varphi(v))$ is linearly independent, meaning $\varphi(v) \neq 0$. Thus $\ker(\varphi) = \{0\}$, so φ is injective.

6.2 Duality

The *dual space* of a vector space V consists of all linear maps from V to the base field k—these are called *linear functionals*. While this might seem like a minor variation on Hom-spaces, duality reveals useful structure: every basis has a "dual basis," every linear map has a "dual map," and the matrix of the dual map is the transpose of the original. In more advanced contexts (differential geometry, functional analysis, representation theory), duality becomes indispensable.

6.2.1 Linear Functionals and the Dual Space

Definition 6.2.1 (Linear functional)

A ***linear functional*** on a vector space V over a field k is a linear map $\ell : V \to k$. In other words, a linear functional is an element of $\mathrm{Hom}(V, k)$.

Example 6.2.2 (Examples of linear functionals)

(a) **Coordinate projections:** On k^n, define $\pi_j(x_1, \ldots, x_n) = x_j$. Each π_j is a linear functional.

(b) **Weighted sums:** Fix $(c_1, \dots, c_n) \in k^n$. Define $\ell(x_1, \dots, x_n) = c_1x_1 + \cdots + c_nx_n$. This is a linear functional on k^n, and *every* linear functional ℓ on k^n has this form, with $c_j = \ell(e_j)$: indeed $\ell(x) = \ell(\sum x_j e_j) = \sum x_j \ell(e_j)$.

(c) **Evaluation:** On $\mathbb{R}[x]$, define $\mathrm{ev}_a(p) = p(a)$ for fixed $a \in \mathbb{R}$. This is a linear functional.

(d) **Integration:** On $\mathbb{R}[x]$, define $\ell(p) = \int_0^1 p(x)\,dx$. Linearity of the integral makes this a linear functional.

(e) **Trace:** On $\mathcal{M}_{n\times n}(k)$, the trace $\mathrm{Tr}(A) = \sum_{i=1}^{n} a_{ii}$ is a linear functional.

Definition 6.2.3 (Dual space)

The ***dual space*** of V, denoted V^*, is the vector space of all linear functionals on V:

$$V^* = \mathrm{Hom}(V, k).$$

This is a vector space under pointwise addition and scalar multiplication: $(\ell_1 + \ell_2)(v) = \ell_1(v) + \ell_2(v)$ and $(\lambda\ell)(v) = \lambda \cdot \ell(v)$.

NOTE Axler writes V' for the dual space and U^0 for the annihilator defined below. In this book the star is reserved for dual spaces; for the dual of a *map* we write φ^t, because φ^* will denote the adjoint of an operator on an inner product space (Chapter 9).

Proposition 6.2.4 (Dimension of the dual space)

If V is finite-dimensional with $\dim(V) = n$, then $\dim(V^*) = n$. Thus $V^* \cong V$ as vector spaces (non-canonically).

Proof. By the formula $\dim \mathrm{Hom}(V, W) = \dim(V) \cdot \dim(W)$ (Corollary 5.2.8):

$$\dim(V^*) = \dim(\mathrm{Hom}(V, k)) = \dim(V) \cdot \dim(k) = n \cdot 1 = n. \qquad \square$$

NOTE The isomorphism $V \cong V^*$ is *non-canonical*—it depends on a choice of basis. There is no "natural" way to identify vectors with linear functionals without extra structure. However, the *double dual* $V^{**} = (V^*)^*$ *does* have a canonical isomorphism with V, given by evaluation: $v \mapsto (\ell \mapsto \ell(v))$ (Theorem 6.2.19).

6.2.2 The Dual Basis

Given a basis of V, there is an associated basis of V^*.

Definition 6.2.5 (Dual basis)

Let $\mathcal{B} = (v_1, \ldots, v_n)$ be a basis of V. The ***dual basis*** is the list $\mathcal{B}^* = (v_1^*, \ldots, v_n^*)$ of linear functionals in V^* defined by

$$v_j^*(v_i) = \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

(A linear map is determined by its values on a basis, and any values may be prescribed; Proposition 5.1.2.) In other words, v_j^* "picks out" the j-th coordinate: if $v = c_1 v_1 + \cdots + c_n v_n$, then $v_j^*(v) = c_j$.

Proposition 6.2.6 (The dual basis is a basis)

The dual basis $(v_1^*, \ldots, v_n^*)$ is a basis of V^*.

Proof. **Linear independence:** Suppose $a_1 v_1^* + \cdots + a_n v_n^* = 0$ (the zero functional). Evaluating at v_i:

$$0 = (a_1 v_1^* + \cdots + a_n v_n^*)(v_i) = a_1 v_1^*(v_i) + \cdots + a_n v_n^*(v_i) = a_i.$$

So all $a_i = 0$.

Spanning: Since we have n linearly independent vectors in V^* and $\dim(V^*) = n$, they form a basis. □

Proposition 6.2.7 (Expansions in a basis and its dual basis)

Let $(v_1, \ldots, v_n)$ be a basis of V with dual basis $(v_1^*, \ldots, v_n^*)$. For any $v \in V$ and any $\ell \in V^*$:

$$v = v_1^*(v)\, v_1 + \cdots + v_n^*(v)\, v_n \qquad \text{and} \qquad \ell = \ell(v_1)\, v_1^* + \cdots + \ell(v_n)\, v_n^*.$$

That is, $v_j^*(v)$ is the j-th coordinate of v in the basis (v_i), and $\ell(v_j)$ is the j-th coordinate of ℓ in the dual basis.

Proof. Write $v = c_1 v_1 + \cdots + c_n v_n$. Applying v_j^*:

$$v_j^*(v) = c_1 v_j^*(v_1) + \cdots + c_n v_j^*(v_n) = c_j.$$

Substituting back gives the first formula. For the second, both sides are linear functionals, and at v_i the right side takes the value $\sum_j \ell(v_j)\, v_j^*(v_i) = \ell(v_i)$. Two linear maps that agree on a basis are equal. □

Example 6.2.8 (Dual basis of the standard basis)

Let $(e_1, \dots, e_n)$ be the standard basis of k^n. The dual basis consists of the coordinate projections:

$$e_j^*(x_1, \dots, x_n) = x_j.$$

If we write elements of k^n as column vectors and represent the functional $x \mapsto c_1x_1 + \cdots + c_nx_n$ by the row vector $(c_1, \dots, c_n)$, so that applying a functional to a vector is the matrix product (row)·(column), then e_j^* is the row vector $(0, \dots, 0, 1, 0, \dots, 0)$ with 1 in position j.

NOTE Each v_j^* depends on the whole basis: changing the other basis vectors can change v_j^* even when v_j stays the same. There is no such thing as "the dual of a vector." In k^2, the basis (e_1, e_2) gives $e_1^*(x_1, x_2) = x_1$, while the basis $(e_1, e_1 + e_2)$, which has the same first vector, gives the first dual functional $(x_1, x_2) \mapsto x_1 - x_2$. This is another way of seeing that the isomorphism $V \cong V^*$ depends on choices.

6.2.3 The Dual Map

Just as a linear map $\varphi : V \to W$ acts on Hom-spaces by composition, it induces a map on dual spaces—but in the *opposite* direction.

Definition 6.2.9 (Dual map (transpose))

Let $\varphi : V \to W$ be a linear map. The ***dual map*** (or ***transpose***) of φ is the map $\varphi^t : W^* \to V^*$ defined by

$$\varphi^t(\ell) = \ell \circ \varphi$$

for each $\ell \in W^*$. In other words, $\varphi^t(\ell)$ is the linear functional on V given by $v \mapsto \ell(\varphi(v))$.

NOTE The dual map goes "backwards": $\varphi : V \to W$ yields $\varphi^t : W^* \to V^*$, because we are precomposing with φ. If you think of linear functionals as

"measurements," then $\varphi^t(\ell)$ is the measurement you get by first applying φ, then measuring with ℓ.

Proposition 6.2.10 (The dual map is linear)
For $\varphi : V \to W$, the dual map $\varphi^t : W^* \to V^*$ is linear.

Proof. For $\ell_1, \ell_2 \in W^*$ and $\lambda \in k$:

$$\varphi^t(\ell_1 + \ell_2) = (\ell_1 + \ell_2) \circ \varphi = \ell_1 \circ \varphi + \ell_2 \circ \varphi = \varphi^t(\ell_1) + \varphi^t(\ell_2),$$
$$\varphi^t(\lambda\ell_1) = (\lambda\ell_1) \circ \varphi = \lambda(\ell_1 \circ \varphi) = \lambda\varphi^t(\ell_1).$$ □

Proposition 6.2.11 (Algebraic properties of the dual map)
Let U, V, W be vector spaces over k.

(a) For $\varphi, \psi : V \to W$, $(\varphi + \psi)^t = \varphi^t + \psi^t$.
(b) For $\varphi : V \to W$ and $\lambda \in k$, $(\lambda\varphi)^t = \lambda\varphi^t$.
(c) For $\varphi : V \to W$ and $\psi : W \to U$, $(\psi \circ \varphi)^t = \varphi^t \circ \psi^t$ (the order reverses).
(d) $(\mathrm{id}_V)^t = \mathrm{id}_{V^*}$.

Proof. (c) is the key property. For $\ell \in U^*$:

$$(\psi \circ \varphi)^t(\ell) = \ell \circ (\psi \circ \varphi) = (\ell \circ \psi) \circ \varphi = \varphi^t(\ell \circ \psi) = \varphi^t(\psi^t(\ell)) = (\varphi^t \circ \psi^t)(\ell).$$

For (a), $(\varphi + \psi)^t(\ell) = \ell \circ (\varphi + \psi) = \ell \circ \varphi + \ell \circ \psi$, because ℓ is additive; (b) is the same computation with $\ell \circ (\lambda\varphi) = \lambda(\ell \circ \varphi)$; and (d) holds because $\ell \circ \mathrm{id}_V = \ell$. □

NOTE Properties (c) and (d) say that $V \mapsto V^*$, $\varphi \mapsto \varphi^t$ reverses the order of composition and preserves identities; in the language of Chapter 8, it is a *contravariant functor* (Definition 8.2.5 and Example 8.2.6). Compare with the transpose of matrices: $(AB)^t = B^tA^t$. This is not a coincidence—the matrix of the dual map *is* the transpose.

6.2.4 Matrix of the Dual Map

The connection between dual maps and matrix transposes explains why the transpose operation is natural.

Theorem 6.2.12 (Matrix of the dual map is the transpose)

Let $\varphi : V \to W$ be a linear map between finite-dimensional spaces. Fix bases $\mathcal{B} = (v_1, \ldots, v_n)$ of V and $\mathcal{C} = (w_1, \ldots, w_m)$ of W, with dual bases $\mathcal{B}^* = (v_1^*, \ldots, v_n^*)$ and $\mathcal{C}^* = (w_1^*, \ldots, w_m^*)$. If

$$A = \mathcal{M}(\varphi, \mathcal{B}, \mathcal{C}) \in \mathcal{M}_{m \times n}(k)$$

is the matrix of φ, then the matrix of φ^t with respect to the dual bases is the transpose:

$$\mathcal{M}(\varphi^t, \mathcal{C}^*, \mathcal{B}^*) = A^t \in \mathcal{M}_{n \times m}(k).$$

Proof. We need to show that $\varphi^t(w_i^*) = \sum_{j=1}^{n} A_{ij} v_j^*$ for each i.

By definition of the matrix A, we have $\varphi(v_j) = \sum_{p=1}^{m} A_{pj} w_p$. Now compute:

$$\varphi^t(w_i^*)(v_j) = w_i^*(\varphi(v_j)) = w_i^* \left(\sum_{p=1}^{m} A_{pj} w_p \right) = \sum_{p=1}^{m} A_{pj}\, w_i^*(w_p) = A_{ij}.$$

By Proposition 6.2.7, $\varphi^t(w_i^*) = \sum_{j=1}^{n} \varphi^t(w_i^*)(v_j)\, v_j^* = \sum_{j=1}^{n} A_{ij} v_j^*$. So the i-th column of $\mathcal{M}(\varphi^t, \mathcal{C}^*, \mathcal{B}^*)$ is the i-th row of A. Hence $\mathcal{M}(\varphi^t, \mathcal{C}^*, \mathcal{B}^*) = A^t$. □

NOTE This theorem explains why the transpose is natural: it is the coordinate representation of the dual map. The reversal of dimensions ($m \times n$ becomes $n \times m$) reflects the reversal of arrows. In terms of row vectors: if $\ell \in W^*$ is represented by the row vector Y and $v \in V$ by the column vector X, then $\ell(\varphi(v)) = YAX$, so φ^t acts on row vectors by $Y \mapsto YA$; writing row vectors as columns turns this into $Y^t \mapsto A^t Y^t$.

6.2.5 Annihilators

Definition 6.2.13 (Annihilator)

Let U be a subspace of V. The ***annihilator*** of U is

$$\operatorname{Ann}(U) = \{\ell \in V^* : \ell(u) = 0 \text{ for all } u \in U\} \subseteq V^*,$$

the set of linear functionals on V that vanish on U.

The annihilator is a subspace of V^*: the zero functional vanishes on U, and if ℓ_1 and ℓ_2 vanish on U, so does $\ell_1 + \lambda\ell_2$. Directly from the definition, $\operatorname{Ann}(\{0\}) = V^*$, $\operatorname{Ann}(V) = \{0\}$, and $U \subseteq U'$ implies $\operatorname{Ann}(U') \subseteq \operatorname{Ann}(U)$: bigger subspaces have smaller annihilators.

Theorem 6.2.14 (Dimension of the annihilator)
Let V be finite-dimensional and $U \subseteq V$ a subspace. The restriction map

$$\rho : V^* \to U^*, \qquad \rho(\ell) = \ell|_U,$$

is linear and surjective, and its kernel is $\mathrm{Ann}(U)$. Thus

$$0 \longrightarrow \mathrm{Ann}(U) \longrightarrow V^* \xrightarrow{\rho} U^* \longrightarrow 0$$

is a short exact sequence (Definition 3.5.1): the inclusion of $\mathrm{Ann}(U)$ is injective, ρ is surjective, and the image of each map is the kernel of the next. Consequently

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

Proof. Linearity of ρ is clear, and $\rho(\ell) = 0$ means exactly that ℓ vanishes on U, so $\ker(\rho) = \mathrm{Ann}(U)$. For surjectivity, let $\psi \in U^*$. Choose a basis $u_1, \dots, u_r$ of U and extend it to a basis $u_1, \dots, u_r, u_{r+1}, \dots, u_n$ of V. Let $\ell \in V^*$ be the functional with $\ell(u_i) = \psi(u_i)$ for $i \leqslant r$ and $\ell(u_i) = 0$ for $i > r$. Then $\ell|_U$ and ψ agree on a basis of U, so $\rho(\ell) = \psi$.

By Rank–Nullity applied to ρ,

$$\dim(V^*) = \dim(\mathrm{Ann}(U)) + \dim(U^*).$$

Since $\dim(V^*) = \dim(V)$ and $\dim(U^*) = \dim(U)$, we get $\dim(\mathrm{Ann}(U)) = \dim(V) - \dim(U)$. □

NOTE The annihilator gives an inclusion-reversing correspondence between subspaces of V and subspaces of V^*: a subspace of dimension r in an n-dimensional V has an annihilator of dimension $n - r$. In Section 7.1, once quotient spaces are available, we will see that $\mathrm{Ann}(U) \cong (V/U)^*$ and $V^*/\mathrm{Ann}(U) \cong U^*$ (Theorem 7.1.10).

6.2.6 Kernel and Image of the Dual Map

The dual map interchanges the roles of kernel and image in a precise way.

Theorem 6.2.15 (Kernel and image of the dual map)
Let $\varphi : V \to W$ be a linear map.

(a) $\ker(\varphi^t) = \mathrm{Ann}(\mathrm{Im}(\varphi))$.

(b) If V and W are finite-dimensional, then $\mathrm{Im}(\varphi^t) = \mathrm{Ann}(\ker(\varphi))$, and $\mathrm{rank}(\varphi^t) = \mathrm{rank}(\varphi)$.

Proof. (a) For $\ell \in W^*$:

$$\ell \in \ker(\varphi^t) \iff \ell \circ \varphi = 0 \iff \ell(\varphi(v)) = 0 \text{ for all } v \in V \iff \ell \in \operatorname{Ann}(\operatorname{Im}(\varphi)).$$

(b) First, $\operatorname{Im}(\varphi^t) \subseteq \operatorname{Ann}(\ker(\varphi))$: if $\ell' = \varphi^t(\ell) = \ell \circ \varphi$, then $\ell'(v) = \ell(\varphi(v)) = \ell(0) = 0$ for every $v \in \ker(\varphi)$. Next we compare dimensions. Let $n = \dim(V)$, $m = \dim(W)$, and $r = \operatorname{rank}(\varphi)$. By Rank–Nullity for φ^t, by (a), and by Theorem 6.2.14,

$$\operatorname{rank}(\varphi^t) = \dim(W^*) - \dim(\ker(\varphi^t)) = m - \dim(\operatorname{Ann}(\operatorname{Im}(\varphi))) = m - (m - r) = r.$$

On the other hand, $\dim(\operatorname{Ann}(\ker(\varphi))) = n - \dim(\ker(\varphi)) = r$, by Theorem 6.2.14 and Rank–Nullity for φ. A subspace of dimension r contained in a space of dimension r is the whole space, so $\operatorname{Im}(\varphi^t) = \operatorname{Ann}(\ker(\varphi))$. □

Corollary 6.2.16 (Injectivity and surjectivity of the dual map)

Let $\varphi : V \to W$ be a linear map between finite-dimensional spaces. Then

(a) φ is surjective if and only if φ^t is injective;

(b) φ is injective if and only if φ^t is surjective.

Proof. (a) By Theorem 6.2.15(a) and Theorem 6.2.14, $\dim(\ker(\varphi^t)) = \dim(W) - \operatorname{rank}(\varphi)$, which is 0 exactly when $\operatorname{rank}(\varphi) = \dim(W)$, i.e., when φ is surjective.

(b) By Theorem 6.2.15(b), $\operatorname{rank}(\varphi^t) = \operatorname{rank}(\varphi) = \dim(V) - \dim(\ker(\varphi))$. This equals $\dim(V^*) = \dim(V)$, i.e., φ^t is surjective, exactly when $\ker(\varphi) = \{0\}$. □

NOTE The implication "φ surjective $\Rightarrow$ φ^t injective" holds without any finiteness assumption: if φ is surjective and $\ell \circ \varphi = 0$, then ℓ vanishes on $\operatorname{Im}(\varphi) = W$, so $\ell = 0$. The other implications use dimension counts.

Corollary 6.2.17 (Row rank equals column rank)

For any matrix $A \in M_{m \times n}(k)$, $\operatorname{rank}(A^t) = \operatorname{rank}(A)$. In words: the largest number of linearly independent rows of A equals the largest number of linearly independent columns.

Proof. The matrix of $L_A : k^n \to k^m$ in the standard bases is A, so by Theorem 6.2.12 the matrix of $(L_A)^t : (k^m)^* \to (k^n)^*$ in the dual bases is A^t. The coordinate isomorphisms $(k^m)^* \cong k^m$ and $(k^n)^* \cong k^n$ given by the dual bases carry $(L_A)^t$ to L_{A^t}, and isomorphisms preserve the dimension of images. Hence $\operatorname{rank}(A^t) = \operatorname{rank}((L_A)^t) = \operatorname{rank}(L_A) = \operatorname{rank}(A)$, using Theorem 6.2.15(b). The rows of A are the columns of A^t, which gives the second sentence. □

6.2.7 The Double Dual

We have seen that $V \cong V^*$ for finite-dimensional V, but this isomorphism required a choice of basis. The double dual $V^{**} = (V^*)^*$ enjoys a more canonical relationship with V.

Definition 6.2.18 (Evaluation map)

The ***evaluation map*** is $\mathrm{ev} : V \to V^{**}$ defined by

$$\mathrm{ev}(v)(\ell) = \ell(v)$$

for $v \in V$ and $\ell \in V^*$. That is, $\mathrm{ev}(v)$ is the linear functional on V^* that "evaluates at v." (It is linear in ℓ because the operations on V^* are pointwise: $\mathrm{ev}(v)(\ell_1 + \lambda\ell_2) = \ell_1(v) + \lambda\ell_2(v)$.)

Theorem 6.2.19 (The double dual)

The evaluation map $\mathrm{ev} : V \to V^{**}$ is linear. If V is finite-dimensional, then ev is an isomorphism.

Proof. **Linearity:** For $v, w \in V$, $\lambda \in k$, and $\ell \in V^*$:

$$\mathrm{ev}(v + \lambda w)(\ell) = \ell(v + \lambda w) = \ell(v) + \lambda\ell(w) = \mathrm{ev}(v)(\ell) + \lambda\,\mathrm{ev}(w)(\ell).$$

Injectivity (for finite-dimensional V): Suppose $\mathrm{ev}(v) = 0$, meaning $\ell(v) = 0$ for all $\ell \in V^*$. If $v \neq 0$, extend $\{v\}$ to a basis $(v, v_2, \ldots, v_n)$ of V; the first dual basis functional ℓ satisfies $\ell(v) = 1 \neq 0$, a contradiction. So $v = 0$.

Isomorphism: For finite-dimensional V, $\dim(V^{**}) = \dim(V^*) = \dim(V)$. Since ev is injective, it is an isomorphism (Corollary 6.1.16). □

NOTE The point is that ev involves no choices. Given a basis $(e_1, \ldots, e_n)$ of V, its dual basis $(e_1^*, \ldots, e_n^*)$ of V^*, and the dual basis $(e_1^{**}, \ldots, e_n^{**})$ of V^{**}, we have $\mathrm{ev}(e_i)(e_j^*) = e_j^*(e_i) = \delta_{ij} = e_i^{**}(e_j^*)$, so $\mathrm{ev}(e_i) = e_i^{**}$. The basis-dependent isomorphisms $V \cong V^*$ and $V^* \cong V^{**}$ compose to an isomorphism that does not depend on the basis.
Moreover, ev is compatible with every linear map $\varphi : V \to W$: writing ev_V and ev_W for the evaluation maps of V and W,

$$(\varphi^t)^t \circ \mathrm{ev}_V = \mathrm{ev}_W \circ \varphi,$$

since for $\ell \in W^*$ both sides send v to the functional $\ell \mapsto \ell(\varphi(v))$. Chapter 8 expresses this by saying that ev is a *natural transformation* (Definition 8.2.9).

NOTE For infinite-dimensional V, the situation is strikingly different. Suppose V has a basis $\{e_i\}_{i\in I}$ indexed by an infinite set I (every vector space has a basis, a consequence of Zorn's lemma). Every element of V is a *finite* linear combination $\sum_{i\in I} x_i e_i$ (with only finitely many $x_i \neq 0$). However, elements of V^* can be arbitrary on the basis: given any function $a : I \to k$, the map $\ell_a(\sum x_i e_i) = \sum x_i a_i$ is a well-defined linear functional. Thus $V^* \cong k^I = \prod_{i\in I} k$, the full product, which is much larger than the direct sum $V \cong \bigoplus_{i\in I} k$; in particular the coordinate functionals e_i^* do not span V^*. The evaluation map $\mathrm{ev} : V \to V^{**}$ is still injective (for $v \neq 0$, some coordinate functional e_i^* has $e_i^*(v) \neq 0$), but it is not surjective.

Question 6.2.20

(a) Let $V = \mathbb{R}^3$ with standard basis (e_1, e_2, e_3). Write down the dual basis explicitly as functions $\mathbb{R}^3 \to \mathbb{R}$.

(b) Let $\varphi : \mathbb{R}^3 \to \mathbb{R}^2$ be defined by $\varphi(x, y, z) = (x + y, y + z)$. Find the matrix of φ with respect to the standard bases, then find the matrix of φ^t with respect to the dual bases.

(c) Let $U = \mathrm{Span}\{(1, 1, 0), (0, 1, 1)\} \subseteq \mathbb{R}^3$. Find a basis of $\mathrm{Ann}(U)$ and check the dimension formula of Theorem 6.2.14.

(d) Let V be finite-dimensional and $U \subseteq V$ a subspace. Prove that $\mathrm{ev}(U) = \mathrm{Ann}(\mathrm{Ann}(U))$, where $\mathrm{ev} : V \to V^{**}$ is the evaluation map. (So under the identification of V with V^{**}, taking the annihilator twice returns U.)

Solution. **(a)** The dual basis consists of the coordinate projections:

$$e_1^*(x, y, z) = x, \quad e_2^*(x, y, z) = y, \quad e_3^*(x, y, z) = z.$$

As row vectors these are $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$.
(b) The matrix of φ is:

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

The matrix of φ^t is the transpose:

$$A^t = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Explicitly, if f_1^*, f_2^* is the dual basis of the standard basis f_1, f_2 of $\mathbb{R}^2$ and e_1^*, e_2^*, e_3^* is that of $\mathbb{R}^3$, then $\varphi^t(f_1^*) = f_1^* \circ \varphi$ is $(x, y, z) \mapsto x + y$, and similarly for f_2^*:

$$\varphi^t(f_1^*) = e_1^* + e_2^*, \quad \varphi^t(f_2^*) = e_2^* + e_3^*.$$

The coefficients form the columns of A^t, as the theorem predicts.

(c) A functional $\ell(x, y, z) = ax + by + cz$ lies in $\mathrm{Ann}(U)$ if and only if it vanishes on the two spanning vectors: $a + b = 0$ and $b + c = 0$. So $b = -a$ and $c = a$, and $\mathrm{Ann}(U)$ is spanned by the single functional

$$\ell(x, y, z) = x - y + z, \qquad \text{i.e., } \ell = e_1^* - e_2^* + e_3^*.$$

The two spanning vectors of U are linearly independent, so $\dim(U) = 2$, and indeed $\dim(\mathrm{Ann}(U)) = 1 = 3 - 2$.

(d) Here $\mathrm{Ann}(U) \subseteq V^*$ and $\mathrm{Ann}(\mathrm{Ann}(U)) = \{\xi \in V^{**} : \xi(\ell) = 0 \text{ for all } \ell \in \mathrm{Ann}(U)\}$. If $u \in U$ and $\ell \in \mathrm{Ann}(U)$, then $\mathrm{ev}(u)(\ell) = \ell(u) = 0$; hence $\mathrm{ev}(U) \subseteq \mathrm{Ann}(\mathrm{Ann}(U))$. Let $n = \dim(V)$. Applying Theorem 6.2.14 twice (to $U \subseteq V$ and to $\mathrm{Ann}(U) \subseteq V^*$),

$$\begin{aligned} \dim(\mathrm{Ann}(\mathrm{Ann}(U))) &= \dim(V^*) - \dim(\mathrm{Ann}(U)) \\ &= n - (n - \dim(U)) = \dim(U). \end{aligned}$$

Since ev is injective, $\dim(\mathrm{ev}(U)) = \dim(U)$ as well. A subspace contained in another of the same finite dimension equals it, so $\mathrm{ev}(U) = \mathrm{Ann}(\mathrm{Ann}(U))$.

CHAPTER 7

Quotient Spaces, Eigenvalues, and Jordan Form

7.1 Quotient Spaces

Quotient spaces provide a systematic way to "collapse" a subspace to zero, creating a smaller vector space that sees V only up to vectors of the subspace. This construction is fundamental in linear algebra and pervades abstract algebra.

7.1.1 Construction of the Quotient Space

Definition 7.1.1 (Cosets and quotient space)

Let V be a vector space over k and $U \subseteq V$ a subspace. For $v \in V$, the ***coset*** of v modulo U is

$$v + U = \{v + u : u \in U\}.$$

The ***quotient space*** is the set of all cosets:

$$V/U = \{v + U : v \in V\}.$$

NOTE Two cosets $v + U$ and $w + U$ are equal if and only if $v - w \in U$. Equivalently, the cosets are the equivalence classes of the relation $v \sim w \Leftrightarrow v - w \in U$, and V/U is the quotient set $V/\sim$ of Definition 1.2.7. As a group, V/U is the quotient of the abelian group $(V, +)$ by its subgroup U (Chapter 3); what is new is the scalar multiplication.

Proposition 7.1.2 (Vector space structure on the quotient)

The quotient V/U becomes a vector space over k under the operations:

(a) **Addition:** $(v + U) + (w + U) = (v + w) + U$.

(b) **Scalar multiplication:** $\lambda(v + U) = (\lambda v) + U$.

The zero element is $0 + U = U$, and $-(v + U) = (-v) + U$.

Proof. The key is *well-definedness*: if $v + U = v' + U$ and $w + U = w' + U$, then $v - v' \in U$ and $w - w' \in U$, so $(v + w) - (v' + w') = (v - v') + (w - w') \in U$, meaning $(v + w) + U = (v' + w') + U$.

Similarly for scalar multiplication: if $v + U = v' + U$, then $v - v' \in U$, so $\lambda(v - v') = \lambda v - \lambda v' \in U$, giving $\lambda v + U = \lambda v' + U$.

Each vector space axiom for V/U follows from the same axiom in V, because the operations are computed on representatives. □

7.1.2 The Quotient Map and Exact Sequences

Definition 7.1.3 (Quotient map)

The ***quotient map*** (or ***canonical projection***) is $q : V \to V/U$ defined by $q(v) = v + U$.

Proposition 7.1.4 (Properties of the quotient map)

The quotient map $q : V \to V/U$ satisfies:

(a) q is linear.

(b) q is surjective.

(c) $\ker(q) = U$.

Proof. (a) $q(v + \lambda w) = (v + \lambda w) + U = (v + U) + \lambda(w + U) = q(v) + \lambda q(w)$.

(b) Every coset $v + U$ is $q(v)$.

(c) $v \in \ker(q) \Leftrightarrow v + U = 0 + U \Leftrightarrow v \in U$. □

NOTE This gives a ***short exact sequence*** (Definition 3.5.1):

$$0 \longrightarrow U \xrightarrow{\iota} V \xrightarrow{q} V/U \longrightarrow 0$$

where $\iota : U \hookrightarrow V$ is the inclusion. "Exact" means: ι is injective, $\mathrm{Im}(\iota) = \ker(q)$, and q is surjective. Short exact sequences are a central organizing principle in algebra.

Theorem 7.1.5 (Dimension of the quotient space)

If V is finite-dimensional and $U \subseteq V$ is a subspace, then

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

Proof. Apply the Rank–Nullity Theorem (Theorem 6.1.12) to $q : V \to V/U$:

$$\dim(V) = \dim(\ker(q)) + \dim(\mathrm{Im}(q)) = \dim(U) + \dim(V/U). \qquad \square$$

Example 7.1.6 (A concrete quotient)

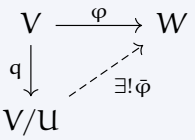
Let $V = \mathbb{R}^3$ and $U = \mathrm{Span}\{(1,0,0)\}$, the x-axis. Then $V/U \cong \mathbb{R}^2$:
The coset $(a,b,c) + U$ consists of all points $\{(a+t,b,c) : t \in \mathbb{R}\}$—the line through (a,b,c) parallel to the x-axis. Two points have the same coset if and only if they have the same y and z coordinates. The map $(a,b,c)+U \mapsto (b,c)$ is an isomorphism $V/U \xrightarrow{\sim} \mathbb{R}^2$.

7.1.3 Universal Property of the Quotient

The quotient space is characterized by a universal property that explains its importance.

Theorem 7.1.7 (Universal property of the quotient)

Let $\varphi : V \to W$ be a linear map with $U \subseteq \ker(\varphi)$. Then there exists a unique linear map $\bar{\varphi} : V/U \to W$ such that $\varphi = \bar{\varphi} \circ q$:



Moreover, $\bar{\varphi}$ is given by $\bar{\varphi}(v+U) = \varphi(v)$.

Proof. **Well-defined:** If $v+U = v'+U$, then $v-v' \in U \subseteq \ker(\varphi)$, so $\varphi(v) = \varphi(v')$.

Linearity: $\bar{\varphi}((v+U)+\lambda(w+U)) = \bar{\varphi}((v+\lambda w)+U) = \varphi(v+\lambda w) = \varphi(v) + \lambda\varphi(w) = \bar{\varphi}(v+U) + \lambda\bar{\varphi}(w+U)$.

Uniqueness: Any $\psi : V/U \to W$ with $\varphi = \psi \circ q$ must satisfy $\psi(v+U) = \psi(q(v)) = \varphi(v) = \bar{\varphi}(v+U)$. $\square$

Corollary 7.1.8 (First Isomorphism Theorem)

For any linear map $\varphi : V \to W$:

$$V/\ker(\varphi) \cong \mathrm{Im}(\varphi).$$

The isomorphism is $\bar{\varphi} : v + \ker(\varphi) \mapsto \varphi(v)$.

Proof. By the universal property with $U = \ker(\varphi)$, the map $\bar{\varphi} : V/\ker(\varphi) \to W$ is well-defined and linear. Its image is exactly $\mathrm{Im}(\varphi)$. Injectivity: if $\bar{\varphi}(v + \ker(\varphi)) = 0$, then $\varphi(v) = 0$, so $v \in \ker(\varphi)$, meaning $v + \ker(\varphi) = 0 + \ker(\varphi)$. □

NOTE This theorem is remarkably useful. It says that every surjective linear map $\varphi : V \twoheadrightarrow W$ is "really" a quotient map: $V \to V/\ker(\varphi) \cong W$. Dually, every injective map $\iota : U \hookrightarrow V$ makes U isomorphic to a subspace of V.

7.1.4 Correspondence Between Subspaces

The quotient construction establishes a bijection between certain subspaces.

Theorem 7.1.9 (Correspondence theorem for subspaces)

Let $U \subseteq V$ be a subspace and $q : V \to V/U$ the quotient map. There is a bijection

$$\{\text{subspaces } W \subseteq V \text{ with } U \subseteq W\} \longleftrightarrow \{\text{subspaces of } V/U\}$$

given by:

- Forward: $W \mapsto W/U = \{w + U : w \in W\} = q(W)$.
- Backward: $\bar{W} \mapsto q^{-1}(\bar{W}) = \{v \in V : v + U \in \bar{W}\}$.

Moreover, this bijection preserves inclusions: $W_1 \subseteq W_2 \Leftrightarrow W_1/U \subseteq W_2/U$.

Proof. Both maps produce subspaces, because the image and the preimage of a subspace under a linear map are subspaces. We verify that the maps are inverse to each other.

Forward then backward: Starting with $W \supseteq U$, we get W/U, then $q^{-1}(W/U) = \{v \in V : v + U \in W/U\}$. But $v + U \in W/U$ iff $v + U = w + U$ for some $w \in W$, iff $v - w \in U \subseteq W$ for some $w \in W$, iff $v \in W$. So $q^{-1}(W/U) = W$.

Backward then forward: Starting with $\bar{W} \subseteq V/U$, let $W = q^{-1}(\bar{W})$. Note $U = q^{-1}(0) \subseteq q^{-1}(\bar{W}) = W$. Then $W/U = q(W) = q(q^{-1}(\bar{W})) = \bar{W}$ (surjectivity of q ensures $q(q^{-1}(S)) = S$ for any $S \subseteq V/U$).

Finally, both maps preserve inclusions, since images and preimages do; and if $W_1/U \subseteq W_2/U$, applying the backward map gives $W_1 = q^{-1}(W_1/U) \subseteq q^{-1}(W_2/U) = W_2$. □

NOTE For instance, the maximal proper subspaces of V containing U correspond to the maximal proper subspaces of V/U. This is the vector-space version of the Correspondence Theorem for groups (Theorem 3.3.7).

7.1.5 Quotients and Duality

Quotient spaces give a second description of the annihilator $\mathrm{Ann}(U)$ of Definition 6.2.13.

Theorem 7.1.10 (Annihilators and quotients)

Let $U \subseteq V$ be a subspace and $q : V \to V/U$ the quotient map.

(a) The dual map $q^t : (V/U)^* \to V^*$, $\psi \mapsto \psi \circ q$, is injective with image $\mathrm{Ann}(U)$. Hence $\mathrm{Ann}(U) \cong (V/U)^*$.

(b) If V is finite-dimensional, the restriction map $\rho : V^* \to U^*$, $\ell \mapsto \ell|_U$, induces an isomorphism $V^*/\mathrm{Ann}(U) \cong U^*$.

Proof. (a) If $\psi \circ q = 0$, then ψ vanishes on $\mathrm{Im}(q) = V/U$, so $\psi = 0$. Every $\psi \circ q$ vanishes on $U = \ker(q)$, so $\mathrm{Im}(q^t) \subseteq \mathrm{Ann}(U)$. Conversely, if $\ell \in \mathrm{Ann}(U)$, then $U \subseteq \ker(\ell)$, so by the universal property (Theorem 7.1.7) $\ell = \bar{\ell} \circ q = q^t(\bar{\ell})$ for some $\bar{\ell} \in (V/U)^*$.

(b) By Theorem 6.2.14, ρ is surjective with kernel $\mathrm{Ann}(U)$. Apply the First Isomorphism Theorem (Corollary 7.1.8). □

NOTE For finite-dimensional V, part (a) recovers $\dim(\mathrm{Ann}(U)) = \dim(V/U) = \dim(V) - \dim(U)$, in agreement with Theorem 6.2.14. Note that part (a) needs no finiteness assumption.

7.2 Polynomials in an Operator

When the domain and codomain of a linear map coincide, we gain additional algebraic structure: the ability to compose maps with themselves. Recall from Chapter 5 that a linear map $\varphi : V \to V$ is a *linear operator*, or *endomorphism*, of V, and that $\mathrm{End}(V) = \mathrm{Hom}(V, V)$ is a ring under addition and composition, with identity id_V (Proposition 5.1.5). Composition is also compatible with scalars, $(\lambda\varphi)\circ\psi = \lambda(\varphi\circ\psi) = \varphi\circ(\lambda\psi)$; one says that $\mathrm{End}(V)$ is a k-*algebra*. For $\dim V = n$,

choosing a basis gives a ring isomorphism $\mathrm{End}(V) \cong \mathcal{M}_{n\times n}(k)$ (Corollary 5.2.10), and the ring is noncommutative once $n \geqslant 2$.

Since endomorphisms can be composed, we can form powers and evaluate polynomials.

Definition 7.2.1 (Powers and polynomials of operators)

For $\varphi \in \mathrm{End}(V)$ and $n \in \mathbb{Z}_{\geqslant 0}$:

$$\varphi^0 = \mathrm{id}_V, \quad \varphi^n = \underbrace{\varphi \circ \varphi \circ \cdots \circ \varphi}_{n \text{ times}}.$$

For a polynomial $p(x) = a_0 + a_1 x + \cdots + a_n x^n \in k[x]$, define

$$p(\varphi) = a_0 \mathrm{id}_V + a_1 \varphi + \cdots + a_n \varphi^n \in \mathrm{End}(V).$$

For $\lambda \in k$ we often write $\varphi - \lambda$ for $\varphi - \lambda \mathrm{id}_V$.

Proposition 7.2.2 (Evaluation homomorphism)

The map $\mathrm{ev}_\varphi : k[x] \to \mathrm{End}(V)$ defined by $p(x) \mapsto p(\varphi)$ is a ring homomorphism. That is:

(a) $(p+q)(\varphi) = p(\varphi) + q(\varphi)$.

(b) $(pq)(\varphi) = p(\varphi) \circ q(\varphi)$.

(c) $1(\varphi) = \mathrm{id}_V$.

In particular, $p(\varphi) \circ q(\varphi) = q(\varphi) \circ p(\varphi)$ for all $p, q \in k[x]$: polynomials in the same operator commute.

Proof. (a) and (c) are immediate. For (b), let $p(x) = \sum a_i x^i$ and $q(x) = \sum b_j x^j$. Then $(pq)(x) = \sum_s c_s x^s$ where $c_s = \sum_{i+j=s} a_i b_j$. Since $\varphi^i \circ \varphi^j = \varphi^{i+j}$ and composition is bilinear,

$$\begin{aligned}(pq)(\varphi) = \sum_s c_s \varphi^s &= \sum_s \Big(\sum_{i+j=s} a_i b_j \Big) \varphi^s \\ &= \Big(\sum_i a_i \varphi^i \Big) \circ \Big(\sum_j b_j \varphi^j \Big) = p(\varphi) \circ q(\varphi).\end{aligned}$$

The last sentence follows from (b) because $pq = qp$ in $k[x]$. □

NOTE A factorization of a polynomial therefore gives a factorization of the operator: if $p(x) = c(x-\lambda_1)\cdots(x-\lambda_d)$, then $p(\varphi) = c(\varphi-\lambda_1)\circ\cdots\circ(\varphi-\lambda_d)$,

with the factors in any order. In the language of Chapter 10, the rule $p \cdot v = p(\varphi)(v)$ makes V a module over the ring $k[x]$. The polynomials p with $p(\varphi) = 0$ form the kernel of ev_φ; in Section 7.9 we show that they are exactly the multiples of a single monic polynomial, the *minimal polynomial* of φ.

7.3 Invariant Subspaces

Understanding an operator often begins with finding subspaces on which its behavior is simpler.

Definition 7.3.1 (Invariant subspace)

A subspace $W \subseteq V$ is ***invariant*** under $\varphi \in \mathrm{End}(V)$ (or ***φ-invariant***) if

$$\varphi(W) \subseteq W,$$

i.e., $\varphi(w) \in W$ for all $w \in W$. In that case the ***restriction*** $\varphi|_W : W \to W$, $w \mapsto \varphi(w)$, is a linear operator on W.

Example 7.3.2 (Basic examples of invariant subspaces)

For any $\varphi \in \mathrm{End}(V)$:

- $\{0\}$ and V are always φ-invariant.
- $\ker(\varphi)$ is φ-invariant: if $\varphi(v) = 0$, then $\varphi(\varphi(v)) = \varphi(0) = 0$.
- $\mathrm{Im}(\varphi)$ is φ-invariant: $\varphi(\mathrm{Im}(\varphi)) = \mathrm{Im}(\varphi^2) \subseteq \mathrm{Im}(\varphi)$.
- More generally, $\ker(\varphi^n)$ and $\mathrm{Im}(\varphi^n)$ are φ-invariant for all $n \geqslant 0$ (a special case of the next proposition).

Proposition 7.3.3 (Kernels and images of commuting operators)

Let $\varphi, \psi \in \mathrm{End}(V)$ with $\varphi \circ \psi = \psi \circ \varphi$. Then $\ker(\psi)$ and $\mathrm{Im}(\psi)$ are φ-invariant. In particular, $\ker(p(\varphi))$ and $\mathrm{Im}(p(\varphi))$ are φ-invariant for every $p \in k[x]$.

Proof. If $\psi(v) = 0$, then $\psi(\varphi(v)) = \varphi(\psi(v)) = 0$, so $\varphi(v) \in \ker(\psi)$. If $w = \psi(u)$, then $\varphi(w) = \varphi(\psi(u)) = \psi(\varphi(u)) \in \mathrm{Im}(\psi)$. The last sentence follows because $p(\varphi)$ commutes with $\varphi = x(\varphi)$ (Proposition 7.2.2). □

7.3.1 Block Diagonal and Block Triangular Matrices

Invariant subspaces lead to simpler matrix representations.

Theorem 7.3.4 (Direct sum of invariant subspaces)

Suppose $V = V_1 \oplus V_2$ where both V_1 and V_2 are φ-invariant. Choose bases $\mathcal{B}_1 = (e_1, \dots, e_m)$ of V_1 and $\mathcal{B}_2 = (e_{m+1}, \dots, e_n)$ of V_2, which together form a basis $\mathcal{B} = (e_1, \dots, e_n)$ of V. Then the matrix of φ with respect to $\mathcal{B}$ is ***block diagonal***:

$$\mathcal{M}(\varphi, \mathcal{B}) = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$$

where $A_1 = \mathcal{M}(\varphi|_{V_1}, \mathcal{B}_1)$ and $A_2 = \mathcal{M}(\varphi|_{V_2}, \mathcal{B}_2)$. The same holds for a decomposition $V = V_1 \oplus \cdots \oplus V_r$ into any number of invariant subspaces.

Proof. For $i \leqslant m$: $\varphi(e_i) \in V_1$ (by φ-invariance), so $\varphi(e_i)$ is a linear combination of $e_1, \dots, e_m$ only, with the coefficients of $\varphi|_{V_1}(e_i)$. For $i > m$: $\varphi(e_i) \in V_2$, so $\varphi(e_i)$ is a linear combination of $e_{m+1}, \dots, e_n$ only. This gives the block structure. The case of r summands is the same. □

NOTE Conversely, if the matrix of φ in a basis is block diagonal, then the spans of the corresponding groups of basis vectors are invariant subspaces whose direct sum is V. The goal of much of spectral theory is to decompose V into a direct sum of invariant subspaces on which φ acts simply.

When only one of the two subspaces is invariant, we still get a simplification, and the quotient space of Section 7.1 describes the second diagonal block.

Proposition 7.3.5 (Induced operator on a quotient)

Let $W \subseteq V$ be φ-invariant, and let $q : V \to V/W$ be the quotient map.

(a) There is a unique linear map $\bar{\varphi} : V/W \to V/W$ with $\bar{\varphi} \circ q = q \circ \varphi$, namely $\bar{\varphi}(v + W) = \varphi(v) + W$:

$$\begin{array}{ccc} V & \xrightarrow{\varphi} & V \\ {\scriptstyle q}\downarrow & & \downarrow{\scriptstyle q} \\ V/W & \xrightarrow[\bar{\varphi}]{} & V/W \end{array}$$

(b) Let V be finite-dimensional, let $w_1, \dots, w_m$ be a basis of W, and extend it to a basis $\mathcal{B} = (w_1, \dots, w_m, v_{m+1}, \dots, v_n)$ of V. Then $\bar{\mathcal{B}} = (v_{m+1} +$

$W, \dots, v_n + W)$ is a basis of V/W, and

$$\mathcal{M}(\varphi, \mathcal{B}) = \begin{pmatrix} A_1 & * \\ 0 & A_2 \end{pmatrix},$$

where $A_1 = \mathcal{M}(\varphi|_W, (w_1, \dots, w_m))$ and $A_2 = \mathcal{M}(\bar{\varphi}, \bar{\mathcal{B}})$. Such a matrix is called ***block upper triangular***.

Proof. (a) The linear map $q \circ \varphi : V \to V/W$ vanishes on W, because $\varphi(W) \subseteq W = \ker(q)$. By the universal property (Theorem 7.1.7) it factors uniquely as $\bar{\varphi} \circ q$.

(b) The cosets $v_{m+1} + W, \dots, v_n + W$ span V/W, since q is surjective and $q(w_i) = 0$; there are $n - m = \dim(V/W)$ of them (Theorem 7.1.5), so they form a basis. For $j \leqslant m$, $\varphi(w_j) \in W$ is a combination of $w_1, \dots, w_m$, which gives the first m columns. For $j > m$, write $\varphi(v_j) = \sum_{i \leqslant m} a_{ij} w_i + \sum_{i > m} a_{ij} v_i$; applying q gives $\bar{\varphi}(v_j + W) = \sum_{i > m} a_{ij}(v_i + W)$, so the lower right block is $\mathcal{M}(\bar{\varphi}, \bar{\mathcal{B}})$. The upper right block $*$ may be nonzero. □

7.4 Eigenvalues and Eigenvectors

The simplest invariant subspaces are one-dimensional. These give rise to the fundamental notion of eigenvalues.

7.4.1 Definitions and Basic Properties

Definition 7.4.1 (Eigenvector and eigenvalue)

Let $\varphi \in \mathrm{End}(V)$. A nonzero vector $v \in V$ is an ***eigenvector*** of φ if

$$\varphi(v) = \lambda v$$

for some scalar $\lambda \in k$. The scalar λ is called the ***eigenvalue*** corresponding to v, and $\lambda \in k$ is an eigenvalue of φ if it corresponds to some eigenvector.

NOTE The condition $\varphi(v) = \lambda v$ says that φ acts on the line spanned by v as scaling by λ. Equivalently, $\mathrm{Span}\{v\}$ is a one-dimensional φ-invariant subspace. Conversely, a one-dimensional invariant subspace $U = \mathrm{Span}\{v\}$ has $\varphi(v) = \lambda v$ for some λ, so it consists of eigenvectors with the same eigenvalue, together with zero.

Proposition 7.4.2 (Characterization of eigenvalues)

For $\varphi \in \mathrm{End}(V)$ and $\lambda \in k$, the following are equivalent:

(a) λ is an eigenvalue of φ.

(b) $\varphi - \lambda \mathrm{id}_V$ is not injective.

(c) $\ker(\varphi - \lambda \mathrm{id}_V) \neq \{0\}$.

If V is finite-dimensional, they are also equivalent to:

(d) $\varphi - \lambda \mathrm{id}_V$ is not surjective (equivalently, not invertible).

(e) $\det(\varphi - \lambda \mathrm{id}_V) = 0$.

Proof. (a) $\Leftrightarrow$ (b) $\Leftrightarrow$ (c): λ is an eigenvalue iff there exists nonzero v with $\varphi(v) = \lambda v$, iff $(\varphi - \lambda \mathrm{id})(v) = 0$ for some nonzero v, iff $\ker(\varphi - \lambda \mathrm{id}) \neq \{0\}$, iff $\varphi - \lambda \mathrm{id}$ is not injective (Theorem 6.1.4).

(b) $\Leftrightarrow$ (d): For an operator on a finite-dimensional space, injective, surjective and invertible are equivalent (Corollary 6.1.16).

(d) $\Leftrightarrow$ (e): An operator on a finite-dimensional space is invertible if and only if its determinant (Definition 5.4.20) is nonzero (Theorem 5.4.8). □

Definition 7.4.3 (Eigenspace)

For $\lambda \in k$, the ***eigenspace*** of φ for λ is

$$E(\lambda, \varphi) = \ker(\varphi - \lambda \mathrm{id}_V) = \{v \in V : \varphi(v) = \lambda v\}.$$

It is a subspace of V, and it is nonzero exactly when λ is an eigenvalue; its nonzero elements are the eigenvectors with eigenvalue λ (the zero vector belongs to it but is not an eigenvector).

Theorem 7.4.4 (Eigenvectors with distinct eigenvalues are linearly independent)

Let $v_1, \ldots, v_m$ be eigenvectors of φ with distinct eigenvalues $\lambda_1, \ldots, \lambda_m$. Then $v_1, \ldots, v_m$ are linearly independent.

Proof. Induction on m. Base case $m = 1$: a single eigenvector is nonzero, hence linearly independent.

Suppose the result holds for $m - 1$ eigenvectors. Assume $\sum_{i=1}^{m} c_i v_i = 0$. Apply φ:

$$\sum_{i=1}^{m} c_i \lambda_i v_i = 0.$$

Multiply the original equation by λ_m and subtract:

$$\sum_{i=1}^{m-1} c_i(\lambda_i - \lambda_m)v_i = 0.$$

By the induction hypothesis (since $v_1, \ldots, v_{m-1}$ are eigenvectors with distinct eigenvalues $\lambda_1, \ldots, \lambda_{m-1}$), we have $c_i(\lambda_i - \lambda_m) = 0$ for $i < m$. Since $\lambda_i \neq \lambda_m$, we get $c_i = 0$. Substituting back: $c_m v_m = 0$, so $c_m = 0$ (since $v_m \neq 0$). □

Corollary 7.4.5 (Bound on the number of eigenvalues)

A linear operator on an n-dimensional space has at most n distinct eigenvalues.

Proof. If $\lambda_1, \ldots, \lambda_m$ are distinct eigenvalues with eigenvectors $v_1, \ldots, v_m$, then $v_1, \ldots, v_m$ are linearly independent. In an n-dimensional space, any linearly independent set has at most n elements. □

Corollary 7.4.6 (The sum of eigenspaces is direct)

If $\lambda_1, \ldots, \lambda_m$ are distinct eigenvalues of φ, then the sum $E(\lambda_1, \varphi) + \cdots + E(\lambda_m, \varphi)$ is direct:

$$E(\lambda_1, \varphi) + \cdots + E(\lambda_m, \varphi) = E(\lambda_1, \varphi) \oplus \cdots \oplus E(\lambda_m, \varphi).$$

If V is finite-dimensional, then $\sum_i \dim E(\lambda_i, \varphi) \leqslant \dim V$.

Proof. Suppose $u_1 + \cdots + u_m = 0$ with $u_i \in E(\lambda_i, \varphi)$. If some u_i were nonzero, the nonzero ones would be eigenvectors with distinct eigenvalues satisfying a linear relation with all coefficients equal to 1, contradicting Theorem 7.4.4. So every $u_i = 0$, which is what it means for the sum to be direct. The dimension of a direct sum is the sum of the dimensions of the summands (Proposition 4.7.14), and it is at most $\dim V$. □

7.4.2 The Characteristic Polynomial

Eigenvalues are the roots of a polynomial built from the determinant (Section 5.4). The matrix $xI - A$ has entries in the polynomial ring $k[x]$, and determinants of matrices over a commutative ring are defined by the same Leibniz formula and have the same basic properties as over a field (Example 5.4.25 and the discussion before it).

Definition 7.4.7 (Characteristic polynomial)

Let $A \in \mathcal{M}_{n\times n}(k)$. The ***characteristic polynomial*** of A is

$$\chi_A(x) = \det(xI - A),$$

the determinant of the matrix $xI - A \in \mathcal{M}_{n\times n}(k[x])$. For $\varphi \in \mathrm{End}(V)$ with V finite-dimensional, the characteristic polynomial of φ is $\chi_\varphi = \chi_A$, where $A = \mathcal{M}(\varphi, \mathcal{B})$ for a basis $\mathcal{B}$ of V. (For $V = \{0\}$ we set $\chi_\varphi = 1$.)

Proposition 7.4.8 (Basic properties of the characteristic polynomial)

Let $\varphi \in \mathrm{End}(V)$ with $\dim V = n \geqslant 1$, and let $A = \mathcal{M}(\varphi, \mathcal{B})$.

(a) $\chi_A(x)$ is a monic polynomial of degree n in $k[x]$, and

$$\chi_A(x) = x^n - (\operatorname{Tr} A)\, x^{n-1} + \cdots + (-1)^n \det(A).$$

(b) For $\lambda \in k$, the value of χ_A at λ is $\chi_A(\lambda) = \det(\lambda I - A)$.

(c) χ_φ does not depend on the choice of basis $\mathcal{B}$.

(d) The roots of χ_φ in k are exactly the eigenvalues of φ.

Proof. (a) and (b): By the Leibniz formula, $\det(xI - A)$ is a sum, over permutations, of $\pm$ a product of n entries of $xI - A$, one from each row and each column. Each entry is a polynomial in x, so the sum is a polynomial in x; and substituting $x = \lambda$ term by term gives the Leibniz formula for $\det(\lambda I - A)$, which proves (b) (see also Example 5.4.25). The identity permutation contributes $\prod_i (x - a_{ii}) = x^n - (a_{11} + \cdots + a_{nn})x^{n-1} + \cdots$. Every other permutation moves at least two indices, so its product contains at most $n - 2$ diagonal entries and has degree at most $n - 2$. Hence χ_A is monic of degree n with x^{n-1}-coefficient $-\operatorname{Tr} A$. Its constant term is $\chi_A(0) = \det(-A) = (-1)^n \det(A)$ by (b).

(c) If $\mathcal{B}'$ is another basis, then $B = \mathcal{M}(\varphi, \mathcal{B}') = P^{-1}AP$ for an invertible matrix P (Corollary 5.3.4). Then $xI - B = P^{-1}(xI - A)P$ in $\mathcal{M}_{n\times n}(k[x])$, so by multiplicativity of the determinant (Theorem 5.4.7, which holds over the commutative ring $k[x]$),

$$\det(xI - B) = \det(P)^{-1} \det(xI - A) \det(P) = \det(xI - A).$$

(d) By (b), $\chi_\varphi(\lambda) = \det(\lambda \mathrm{id}_V - \varphi) = (-1)^n \det(\varphi - \lambda \mathrm{id}_V)$, which vanishes if and only if λ is an eigenvalue (Proposition 7.4.2). □

Example 7.4.9 (Characteristic polynomial of a 2×2 matrix)

For $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$:

$$\begin{aligned} \chi_A(x) = \det \begin{pmatrix} x-a & -b \\ -c & x-d \end{pmatrix} &= (x-a)(x-d) - bc \\ &= x^2 - (a+d)x + (ad-bc) \\ &= x^2 - (\operatorname{Tr} A)x + \det A. \end{aligned}$$

Definition 7.4.10 (Algebraic and geometric multiplicity)

Let λ be an eigenvalue of $\varphi \in \operatorname{End}(V)$, V finite-dimensional. The ***algebraic multiplicity*** of λ is the largest integer e such that $(x-\lambda)^e$ divides $\chi_\varphi(x)$. The ***geometric multiplicity*** of λ is $\dim E(\lambda, \varphi)$.

When χ_φ is a product of linear factors, the algebraic multiplicity simply counts how often $x - \lambda$ occurs:

Lemma 7.4.11 (Multiplicity of a root in a product of linear factors)

Let $c_1, \ldots, c_n, \lambda \in k$. The largest e such that $(x-\lambda)^e$ divides $\prod_{j=1}^{n}(x-c_j)$ is the number of indices j with $c_j = \lambda$.

Proof. Let e_0 be the number of j with $c_j = \lambda$, and write $\prod_j (x-c_j) = (x-\lambda)^{e_0} h(x)$ with $h = \prod_{c_j \neq \lambda}(x-c_j)$. Then $h(\lambda) = \prod_{c_j \neq \lambda}(\lambda - c_j) \neq 0$. Clearly $(x-\lambda)^{e_0}$ divides the product. If $(x-\lambda)^{e_0+1}$ divided it, say $(x-\lambda)^{e_0} h = (x-\lambda)^{e_0+1} g$, then cancelling $(x-\lambda)^{e_0}$ (possible because $k[x]$ has no zero divisors, Lemma 4.3.3) would give $h = (x-\lambda)g$, hence $h(\lambda) = 0$, a contradiction. □

7.4.3 Diagonalizable Operators

The best-case scenario: V has a basis of eigenvectors.

Definition 7.4.12 (Diagonalizable operator)

An operator $\varphi \in \operatorname{End}(V)$ on a finite-dimensional space is ***diagonalizable*** if there exists a basis of V consisting entirely of eigenvectors of φ.

NOTE If $(v_1, \dots, v_n)$ is a basis of eigenvectors with $\varphi(v_i) = \lambda_i v_i$, then the matrix of φ is

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

The eigenvalues appear on the diagonal (possibly with repetition).

Theorem 7.4.13 (Characterizations of diagonalizability)

Let $\varphi \in \mathrm{End}(V)$ with $\dim(V) = n$, and let $\lambda_1, \dots, \lambda_m$ be the distinct eigenvalues of φ. The following are equivalent:

(a) φ is diagonalizable.

(b) The matrix of φ with respect to some basis is diagonal.

(c) $V = E(\lambda_1, \varphi) \oplus \cdots \oplus E(\lambda_m, \varphi)$.

(d) $\dim E(\lambda_1, \varphi) + \cdots + \dim E(\lambda_m, \varphi) = n$.

Proof. (a) $\Leftrightarrow$ (b): The matrix in the basis $(v_1, \dots, v_n)$ is diagonal exactly when each $\varphi(v_j)$ is a multiple of v_j, i.e., when every v_j is an eigenvector.

(a) $\Rightarrow$ (c): Every vector of a basis of eigenvectors lies in some $E(\lambda_i, \varphi)$, so these eigenspaces span V; their sum is direct by Corollary 7.4.6.

(c) $\Rightarrow$ (d): Take dimensions: $\dim V = \dim(E(\lambda_1, \varphi) \oplus \cdots \oplus E(\lambda_m, \varphi)) = \sum \dim E(\lambda_i, \varphi)$.

(d) $\Rightarrow$ (a): The sum $E(\lambda_1, \varphi) + \cdots + E(\lambda_m, \varphi)$ is direct (Corollary 7.4.6) and has dimension $\sum \dim E(\lambda_i, \varphi) = n = \dim V$, so it equals V. Combining bases of the $E(\lambda_i, \varphi)$ gives a basis of V consisting of eigenvectors. □

Corollary 7.4.14 (Enough distinct eigenvalues implies diagonalizability)

If $\varphi \in \mathrm{End}(V)$ has $\dim(V)$ distinct eigenvalues, then φ is diagonalizable.

Proof. Each eigenspace has dimension at least 1 (it contains an eigenvector). With $n = \dim V$ distinct eigenvalues, $\sum \dim E(\lambda_i, \varphi) \geqslant n$. But this sum is also $\leqslant n$ (Corollary 7.4.6), so equality holds, and criterion (d) above is satisfied. □

7.4.4 Examples: Diagonalizable and Non-Diagonalizable

Example 7.4.15 (Diagonal matrices)

The identity matrix I_n has eigenvalue 1 with eigenspace all of k^n. More generally, a diagonal matrix $\operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ has the standard basis vectors as eigenvectors with eigenvalues $\lambda_1, \ldots, \lambda_n$. Such a matrix is (trivially) diagonalizable—it is already diagonal.

Example 7.4.16 (A non-diagonalizable matrix)

Consider $\varphi : \mathbb{R}^2 \to \mathbb{R}^2$ with matrix

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

The characteristic polynomial is $\chi_A(x) = (x-1)^2$, so $\lambda = 1$ is the only eigenvalue.

The eigenspace: $E(1, \varphi) = \ker(A - I) = \ker \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \operatorname{Span}\{e_1\}$.

This has dimension 1, but $\dim(\mathbb{R}^2) = 2$. So φ is **not diagonalizable**.

Geometrically, φ is a "shear": it fixes the x-axis pointwise and slides other points parallel to the x-axis. The only eigenvector direction is along the x-axis.

Example 7.4.17 (No eigenvectors over $\mathbb{R}$)

Consider rotation by $90°$ in $\mathbb{R}^2$:

$$R = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

The characteristic polynomial is $\chi_R(x) = x^2 + 1$, which has no real roots. So R has *no eigenvalues* over $\mathbb{R}$, hence no eigenvectors. Over $\mathbb{C}$, the eigenvalues are $\pm i$ with eigenvectors $(1, \mp i)^t$, and R is diagonalizable as a complex matrix. This illustrates that diagonalizability depends on the scalar field.

NOTE The rotation example shows that over $\mathbb{R}$ an operator may have no eigenvalues at all; Section 7.10 explains what can be said instead. Even over $\mathbb{C}$, where every polynomial has roots, an operator may fail to be diagonalizable

(like the shear above). Understanding non-diagonalizable operators requires the Jordan normal form, which we develop in Section 7.8.

Question 7.4.18 (Exercises on quotient spaces, operators, and eigenvalues)

(a) Let $V = \mathbb{R}^4$ and $U = \{(x, y, z, w) : x + y = 0, z = w\}$. Find $\dim(U)$, $\dim(V/U)$, and describe V/U explicitly.

(b) Prove that if $\varphi, \psi \in \mathrm{End}(V)$ commute ($\varphi \circ \psi = \psi \circ \varphi$), then every eigenspace of φ is ψ-invariant.

(c) Let $\varphi \in \mathrm{End}(V)$ be diagonalizable with distinct eigenvalues $\lambda_1, \ldots, \lambda_m$. Prove that $p(\varphi) = 0$ where $p(x) = \prod_{i=1}^{m}(x - \lambda_i)$.

(d) Show that the quotient map $q : V \to V/U$ induces a bijection between $\{\text{linear maps } V/U \to W\}$ and $\{\text{linear maps } V \to W \text{ vanishing on } U\}$.

(e) Let $A = \begin{pmatrix} 4 & 2 \\ 1 & 3 \end{pmatrix}$. Find the eigenvalues and eigenspaces of A. Is A diagonalizable? If so, find a matrix P such that $P^{-1}AP$ is diagonal.

Solution. **(a)** The map $\pi : \mathbb{R}^4 \to \mathbb{R}^2$ defined by $\pi(x, y, z, w) = (x + y, z - w)$ is linear and has $\ker(\pi) = U$. It is surjective, since $\pi(a, 0, b, 0) = (a, b)$. By Rank–Nullity, $\dim(U) = 4 - 2 = 2$; a basis is $\{(-1, 1, 0, 0), (0, 0, 1, 1)\}$.
By the dimension formula: $\dim(V/U) = \dim(V) - \dim(U) = 4 - 2 = 2$.
By the First Isomorphism Theorem, $\bar{\pi} : V/U \to \mathbb{R}^2$ is an isomorphism. Explicitly, the coset $(x, y, z, w) + U$ is determined by the pair $(x + y, z - w)$.
(b) Let $v \in E(\lambda, \varphi)$, i.e., $\varphi(v) = \lambda v$. We need $\psi(v) \in E(\lambda, \varphi)$, i.e., $\varphi(\psi(v)) = \lambda\psi(v)$.
Compute: $\varphi(\psi(v)) = \psi(\varphi(v)) = \psi(\lambda v) = \lambda\psi(v)$. ✓
(We used commutativity in the first equality. This is also a special case of Proposition 7.3.3, since ψ commutes with $\varphi - \lambda\mathrm{id}$.)
(c) Since φ is diagonalizable, $V = \bigoplus_{i=1}^{m} E(\lambda_i, \varphi)$. Let $v \in E(\lambda_j, \varphi)$. The factors of $p(\varphi) = (\varphi - \lambda_1) \cdots (\varphi - \lambda_m)$ commute (Proposition 7.2.2), so we may apply $\varphi - \lambda_j$ first:

$$p(\varphi)(v) = \Big(\prod_{i \neq j}(\varphi - \lambda_i)\Big)\big((\varphi - \lambda_j)(v)\big) = 0.$$

Since this holds on each eigenspace and the eigenspaces span V, we have $p(\varphi) = 0$.
(d) Define $\Phi : \mathrm{Hom}(V/U, W) \to \{T \in \mathrm{Hom}(V, W) : U \subseteq \ker(T)\}$ by $\Phi(\bar{T}) = \bar{T} \circ q$. This lands in the right set, because $q(U) = 0$.

Injective: If $\bar{T} \circ q = 0$, then $\bar{T}(v + U) = \bar{T}(q(v)) = 0$ for all v, so $\bar{T} = 0$.
Surjective: Given $T : V \to W$ with $U \subseteq \ker(T)$, Theorem 7.1.7 gives $\bar{T} : V/U \to W$ with $\bar{T}(v + U) = T(v)$, so $\Phi(\bar{T}) = \bar{T} \circ q = T$.
(Both sides are vector spaces and Φ is linear, so Φ is even an isomorphism of vector spaces.)
(e) Characteristic polynomial: $\chi_A(x) = (x-4)(x-3) - 2 = x^2 - 7x + 10 = (x-5)(x-2)$.
Eigenvalues: $\lambda_1 = 5, \lambda_2 = 2$.
Eigenspaces:

- $\lambda = 5$: $(A - 5I)v = 0$ gives $\begin{pmatrix} -1 & 2 \\ 1 & -2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = 0$, so $x = 2y$. Eigenspace: $\operatorname{Span}\{(2,1)^t\}$.
- $\lambda = 2$: $(A - 2I)v = 0$ gives $\begin{pmatrix} 2 & 2 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = 0$, so $x = -y$. Eigenspace: $\operatorname{Span}\{(1,-1)^t\}$.

Since A has $2 = \dim \mathbb{R}^2$ distinct eigenvalues, it is diagonalizable (Corollary 7.4.14).
Let $P = \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix}$ (columns are eigenvectors). Then $P^{-1}AP = \begin{pmatrix} 5 & 0 \\ 0 & 2 \end{pmatrix}$.

7.5 Existence of Eigenvectors and Triangular Form

Over $\mathbb{R}$, the rotation example shows that an operator need not have any eigenvector. Over an algebraically closed field (Chapter 4), such as $\mathbb{C}$, this cannot happen.

Theorem 7.5.1 (Existence of eigenvectors)
Let k be algebraically closed and let $V \neq \{0\}$ be a finite-dimensional vector space over k. Then every $\varphi \in \operatorname{End}(V)$ has an eigenvector: there exist $v \in V$, $v \neq 0$, and $\lambda \in k$ with $\varphi(v) = \lambda v$.

Proof. Let $n = \dim V$ and take any nonzero vector $v \in V$. The $n+1$ vectors v, $\varphi(v), \ldots, \varphi^n(v)$ must be linearly dependent, so there exist $a_0, \ldots, a_n \in k$, not all zero, with

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

Let $p(x) = a_0 + a_1 x + \cdots + a_n x^n$, a nonzero polynomial with $p(\varphi)(v) = 0$. It is not a nonzero constant, since $a_0 v \neq 0$ for $a_0 \neq 0$; so $\deg p = d \geqslant 1$. Since k is algebraically closed, p factors as $p(x) = c(x - \lambda_1)\cdots(x - \lambda_d)$ with $c \neq 0$ and $\lambda_i \in k$ (Proposition 4.3.17). Because evaluation at φ is a ring homomorphism

(Proposition 7.2.2),

$$p(\varphi) = c(\varphi - \lambda_1) \circ \cdots \circ (\varphi - \lambda_d).$$

This operator is not injective, since it kills $v \neq 0$. A composition of injective maps is injective, so some factor $\varphi - \lambda_i$ is not injective. Hence λ_i is an eigenvalue of φ (Proposition 7.4.2). □

NOTE Equivalently, by Proposition 7.4.8(d), χ_φ has a root because it is a nonconstant polynomial over an algebraically closed field. The proof above avoids determinants. Over $\mathbb{R}$, it breaks down exactly where p fails to factor into linear factors; we return to real operators in Section 7.10.

A square matrix $A = (a_{ij})$ is ***upper triangular*** if $a_{ij} = 0$ for $i > j$. If $\mathcal{B} = (v_1, \dots, v_n)$ is a basis of V, the matrix $\mathcal{M}(\varphi, \mathcal{B})$ is upper triangular if and only if $\varphi(v_j) \in \operatorname{Span}(v_1, \dots, v_j)$ for every j, that is, if and only if each subspace $V_j = \operatorname{Span}(v_1, \dots, v_j)$ is φ-invariant.

Theorem 7.5.2 (Upper-triangular form)

Let k be algebraically closed, V a finite-dimensional vector space over k, and $\varphi \in \operatorname{End}(V)$. Then there is a basis of V in which the matrix of φ is upper triangular:

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

Proof. Induction on $n = \dim V$. If $n \leqslant 1$, every matrix of φ is upper triangular.

Let $n \geqslant 2$ and assume the result in dimension $n - 1$. By Theorem 7.5.1, φ has an eigenvector v_1, say $\varphi(v_1) = \lambda_1 v_1$. The line $L = \operatorname{Span}(v_1)$ is φ-invariant, so φ induces $\bar{\varphi} : V/L \to V/L$ with $\bar{\varphi}(q(v)) = q(\varphi(v))$, where $q : V \to V/L$ is the quotient map (Proposition 7.3.5). Intuitively, $\bar{\varphi}$ is "φ modulo v_1." Since $\dim(V/L) = n - 1$, the induction hypothesis gives a basis $u_2, \dots, u_n$ of V/L with $\bar{\varphi}(u_j) \in \operatorname{Span}(u_2, \dots, u_j)$ for each j. Choose $v_j \in V$ with $q(v_j) = u_j$.

The list $(v_1, \dots, v_n)$ spans V: for $v \in V$, write $q(v) = \sum_{j \geqslant 2} c_j u_j$; then $v - \sum_{j \geqslant 2} c_j v_j \in \ker(q) = L = \operatorname{Span}(v_1)$. Having $n = \dim V$ elements, it is a basis. Finally, for $j \geqslant 2$,

$$q(\varphi(v_j)) = \bar{\varphi}(u_j) \in \operatorname{Span}(u_2, \dots, u_j) = q\big(\operatorname{Span}(v_2, \dots, v_j)\big),$$

so $\varphi(v_j) \in \operatorname{Span}(v_2, \dots, v_j) + L = \operatorname{Span}(v_1, \dots, v_j)$. Together with $\varphi(v_1) = \lambda_1 v_1$, this says the matrix of φ in the basis $(v_1, \dots, v_n)$ is upper triangular. (Its first

column is $(\lambda_1, 0, \ldots, 0)^t$, and its lower right $(n-1) \times (n-1)$ block is the matrix of $\bar{\varphi}$ in the basis $(u_2, \ldots, u_n)$, as in Proposition 7.3.5.) □

The diagonal of a triangular matrix carries the information about invertibility and eigenvalues. The next two results hold over any field.

Lemma 7.5.3 (Invertibility from the diagonal)

Let $\varphi \in \mathrm{End}(V)$ have an upper-triangular matrix $A = (a_{ij})$ in the basis $(v_1, \ldots, v_n)$. Then φ is invertible if and only if all the diagonal entries $a_{11}, \ldots, a_{nn}$ are nonzero.

Proof. Write $V_j = \mathrm{Span}(v_1, \ldots, v_j)$, with $V_0 = \{0\}$.

Suppose all $a_{jj} \neq 0$. We show by induction on j that $v_j \in \mathrm{Im}(\varphi)$. Since $\varphi(v_j) = a_{jj}v_j + \sum_{i<j} a_{ij}v_i$,

$$v_j = \frac{1}{a_{jj}}\Big(\varphi(v_j) - \sum_{i<j} a_{ij}v_i\Big),$$

and the right side lies in $\mathrm{Im}(\varphi)$ because $v_1, \ldots, v_{j-1}$ do. So $\mathrm{Im}(\varphi)$ contains a basis, φ is surjective, and hence invertible (Corollary 6.1.16).

Suppose $a_{jj} = 0$ for some j. Then $\varphi(v_j) \in V_{j-1}$, and also $\varphi(V_{j-1}) \subseteq V_{j-1}$, so φ maps the j-dimensional space V_j into the $(j-1)$-dimensional space V_{j-1}. By Corollary 6.1.15, $\varphi|_{V_j}$ is not injective, so φ is not injective and not invertible. □

Corollary 7.5.4 (Eigenvalues from a triangular matrix)

Let $\varphi \in \mathrm{End}(V)$ have an upper-triangular matrix $A = (a_{ij})$ in some basis. Then:

(a) $\lambda \in k$ is an eigenvalue of $\varphi \iff \varphi - \lambda$ is not invertible $\iff \lambda = a_{jj}$ for some j.

(b) $\chi_\varphi(x) = (x - a_{11})(x - a_{22}) \cdots (x - a_{nn})$, and the number of times an eigenvalue λ appears on the diagonal of A is its algebraic multiplicity.

In particular, every upper-triangular matrix representing φ has the same diagonal entries, up to order.

Proof. (a) The first equivalence is Proposition 7.4.2. For the second, apply Lemma 7.5.3 to $\varphi - \lambda$, whose matrix $A - \lambda I$ is upper triangular with diagonal entries $a_{jj} - \lambda$.

(b) The matrix $xI - A$ is upper triangular with diagonal entries $x - a_{jj}$, and the determinant of a triangular matrix, over $k[x]$ as over k, is the product of its diagonal entries (Proposition 5.4.10). The statement about multiplicities is

Lemma 7.4.11. The last sentence follows because χ_φ does not depend on the basis. □

NOTE Together, Theorem 7.5.2 and Corollary 7.5.4 show that over an algebraically closed field χ_φ splits into linear factors, its roots being the diagonal entries of any triangular matrix of φ. We will soon see a more conceptual description of the algebraic multiplicity: it is the dimension of the generalized eigenspace (Theorem 7.6.14).

7.6 Generalized Eigenspaces

The shear $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ has too few eigenvectors to be diagonalizable. To decompose such operators we enlarge the eigenspaces, admitting vectors that are not eigenvectors but are killed by some power of $\varphi - \lambda$. It is simplest to start with the eigenvalue 0, where the relevant operator is φ itself.

7.6.1 Generalized Kernels

Definition 7.6.1 (Generalized kernel)

The ***generalized kernel*** of $\varphi \in \operatorname{End}(V)$ is

$$\operatorname{gker}(\varphi) = \{v \in V : \varphi^m(v) = 0 \text{ for some } m \geqslant 1\}.$$

These are all the vectors that are eventually sent to 0 by repeatedly applying φ.

NOTE The notation $\operatorname{gker}(\varphi)$ is Auroux's; it is not standard.

The difference between the kernel and the generalized kernel can be summed up as follows: a vector is in the kernel if one application of φ sends it to 0, and in the generalized kernel if some number of applications does. If $\varphi^m(v) = 0$, then also $\varphi^{m+1}(v) = 0$, so the kernels of the powers of φ form an increasing chain

$$\{0\} = \ker(\varphi^0) \subseteq \ker(\varphi) \subseteq \ker(\varphi^2) \subseteq \ker(\varphi^3) \subseteq \cdots,$$

and $\operatorname{gker}(\varphi)$ is their union. The union of an increasing chain of subspaces is a subspace, so $\operatorname{gker}(\varphi)$ is a subspace of V.

Proposition 7.6.2 (Stabilization of kernels)

Let $\varphi \in \operatorname{End}(V)$.

(a) If $\ker(\varphi^m) = \ker(\varphi^{m+1})$ for some $m \geqslant 0$, then $\ker(\varphi^{m+j}) = \ker(\varphi^m)$ for all $j \geqslant 0$, and $\operatorname{gker}(\varphi) = \ker(\varphi^m)$.
(b) If V is finite-dimensional with $\dim V = n$, then $\ker(\varphi^n) = \ker(\varphi^{n+1})$, and hence $\operatorname{gker}(\varphi) = \ker(\varphi^n)$.

Proof. (a) We show $\ker(\varphi^{m+2}) = \ker(\varphi^{m+1})$; repeating the argument gives the rest. The inclusion $\supseteq$ always holds. If $\varphi^{m+2}(v) = 0$, then $\varphi^{m+1}(\varphi(v)) = 0$, so $\varphi(v) \in \ker(\varphi^{m+1}) = \ker(\varphi^m)$. Hence $\varphi^{m+1}(v) = \varphi^m(\varphi(v)) = 0$. So the chain is constant from $\ker(\varphi^m)$ on, and its union is $\ker(\varphi^m)$.

(b) Let $d_j = \dim \ker(\varphi^j)$. If $\ker(\varphi^j) = \ker(\varphi^{j+1})$ for some $j < n$, then (a) gives $\ker(\varphi^n) = \ker(\varphi^{n+1})$. Otherwise each inclusion $\ker(\varphi^j) \subsetneq \ker(\varphi^{j+1})$ for $j < n$ is strict, so $0 = d_0 < d_1 < \cdots < d_n$, whence $d_n \geqslant n$ and $\ker(\varphi^n) = V$; then also $\ker(\varphi^{n+1}) = V$. □

Example 7.6.3

Let $\varphi : k^2 \to k^2$ be given by $e_1 \mapsto 0$, $e_2 \mapsto e_1$, with matrix $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. Then $\ker(\varphi) = k \cdot e_1$, but $\ker(\varphi^2) = \operatorname{gker}(\varphi) = k^2$. Note also that $\operatorname{Im}(\varphi) = k \cdot e_1 = \ker(\varphi)$, so $V \neq \ker(\varphi) \oplus \operatorname{Im}(\varphi)$ here.

Definition 7.6.4 (Nilpotent operator)

An operator $\varphi \in \operatorname{End}(V)$ is ***nilpotent*** if $\varphi^m = 0$ for some positive integer m. Then $\operatorname{gker}(\varphi) = V$. When V is finite-dimensional the converse holds, because $\operatorname{gker}(\varphi) = \ker(\varphi^{\dim V})$ by Proposition 7.6.2(b). If $V \neq \{0\}$, the smallest such m is called the ***index of nilpotency*** of φ.

As the example shows, in general $V \neq \ker(\varphi) \oplus \operatorname{Im}(\varphi)$. Replacing φ by a high enough power repairs this:

Lemma 7.6.5 (Kernel–image decomposition for a power)

Let V be finite-dimensional, $\varphi \in \operatorname{End}(V)$, and let $m \geqslant 0$ satisfy $\ker(\varphi^m) = \ker(\varphi^{m+1})$ (for example $m = \dim V$). Then

$$V = \ker(\varphi^m) \oplus \operatorname{Im}(\varphi^m),$$

both summands are φ-invariant, and φ restricts to an isomorphism of $\operatorname{Im}(\varphi^m)$ onto itself. In particular, $V = \operatorname{gker}(\varphi) \oplus \operatorname{Im}(\varphi^{\dim V})$.

Proof. Both summands are φ-invariant by Proposition 7.3.3.

Intersection is trivial: Suppose $v \in \ker(\varphi^m) \cap \operatorname{Im}(\varphi^m)$. Write $v = \varphi^m(u)$ for some $u \in V$. Then $\varphi^{2m}(u) = \varphi^m(v) = 0$, so $u \in \ker(\varphi^{2m}) = \ker(\varphi^m)$ by Proposition 7.6.2(a). Thus $v = \varphi^m(u) = 0$.

Sum is all of V: By Rank–Nullity applied to φ^m, $\dim V = \dim\ker(\varphi^m) + \dim\operatorname{Im}(\varphi^m)$. Since the intersection is trivial, the sum $\ker(\varphi^m) + \operatorname{Im}(\varphi^m)$ is direct and has dimension $\dim V$. Hence it equals V.

Restriction to the image: The kernel of $\varphi|_{\operatorname{Im}(\varphi^m)}$ is $\ker(\varphi) \cap \operatorname{Im}(\varphi^m) \subseteq \ker(\varphi^m) \cap \operatorname{Im}(\varphi^m) = \{0\}$ (here $\ker(\varphi) \subseteq \ker(\varphi^{m+1}) = \ker(\varphi^m)$). An injective operator on a finite-dimensional space is an isomorphism (Corollary 6.1.16).

The last sentence follows from Proposition 7.6.2(b). □

NOTE This decomposition is the key structural result for understanding non-diagonalizable operators. The generalized kernel $\operatorname{gker}(\varphi)$ captures the "nilpotent part" of φ: the restriction of φ to it is nilpotent. On the complementary invariant subspace $\operatorname{Im}(\varphi^m)$, the restriction of φ is an isomorphism.

7.6.2 Generalized Eigenspaces

Now we apply the same idea to $\varphi - \lambda$ for an arbitrary scalar λ. From now on in this section, V is finite-dimensional with $\dim V = n$.

Definition 7.6.6 (Generalized eigenspace)

Let $\varphi \in \operatorname{End}(V)$ and $\lambda \in k$. Set

$$V_\lambda = \operatorname{gker}(\varphi - \lambda) = \{v \in V : (\varphi - \lambda)^m(v) = 0 \text{ for some } m \geqslant 1\} = \ker(\varphi - \lambda)^n,$$

the last equality by Proposition 7.6.2(b). A nonzero vector of V_λ is a ***generalized eigenvector*** of φ with eigenvalue λ. When λ is an eigenvalue of φ, V_λ is called the ***generalized eigenspace*** of φ for λ.

NOTE Every eigenvector with eigenvalue λ is a generalized eigenvector, since $E(\lambda, \varphi) = \ker(\varphi - \lambda) \subseteq V_\lambda$; but V_λ may be strictly larger. Its elements are the vectors that are "eventually killed" by $\varphi - \lambda$.

Example 7.6.7 (Generalized eigenspace of the shear)

Recall the shear $\varphi : \mathbb{R}^2 \to \mathbb{R}^2$ with matrix $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. The only eigenvalue is $\lambda = 1$, and the eigenspace is $E(1, \varphi) = \ker(A - I) = \ker \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \mathrm{Span}\{e_1\}$, which is one-dimensional.
However, $(A - I)^2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}^2 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = 0$.
So the generalized eigenspace is $V_1 = \ker(A - I)^2 = \mathbb{R}^2$. The vector e_2 is a generalized eigenvector that is not an eigenvector: $(A - I)e_2 = e_1 \neq 0$, but $(A - I)^2 e_2 = 0$.

Proposition 7.6.8 (Basic properties of generalized eigenspaces)
Let $\varphi \in \mathrm{End}(V)$ and $\lambda, \mu \in k$. Then:

(a) $E(\lambda, \varphi) \subseteq V_\lambda$, and $V_\lambda \neq \{0\}$ if and only if λ is an eigenvalue of φ.
(b) V_λ is φ-invariant.
(c) The restriction of $\varphi - \lambda$ to V_λ is nilpotent.
(d) If $\mu \neq \lambda$, the restriction of $\varphi - \mu$ to V_λ is invertible. Consequently $V_\lambda \cap V_\mu = \{0\}$, and if $V_\lambda \neq \{0\}$, the only eigenvalue of $\varphi|_{V_\lambda}$ is λ.

Proof. (a) The inclusion is clear. If $v \in V_\lambda$ is nonzero, let $r \geqslant 0$ be the largest integer with $w = (\varphi - \lambda)^r(v) \neq 0$; then $(\varphi - \lambda)(w) = 0$, so w is an eigenvector with eigenvalue λ. Conversely, an eigenvector lies in $E(\lambda, \varphi) \subseteq V_\lambda$.

(b) $V_\lambda = \ker\big((\varphi - \lambda)^n\big)$ and $(\varphi - \lambda)^n$ is a polynomial in φ, so this is Proposition 7.3.3.

(c) By definition $(\varphi - \lambda)^n$ is zero on V_λ, and V_λ is invariant under $\varphi - \lambda$ by (b).

(d) Let $v \in V_\lambda$ with $(\varphi - \mu)(v) = 0$. Then $(\varphi - \lambda)(v) = (\mu - \lambda)v$, so $0 = (\varphi - \lambda)^n(v) = (\mu - \lambda)^n v$. As $\mu - \lambda \neq 0$, we get $v = 0$. Thus $\varphi - \mu$ is injective on the invariant finite-dimensional subspace V_λ, hence invertible there (Corollary 6.1.16). Then $(\varphi - \mu)^n$ is injective on V_λ too; since it kills V_μ, we get $V_\lambda \cap V_\mu = \{0\}$. Finally, an eigenvalue μ of $\varphi|_{V_\lambda}$ is a scalar for which $\varphi - \mu$ is not injective on V_λ, so $\mu = \lambda$. □

Proposition 7.6.9 (Invariant splitting)
Let $\varphi \in \mathrm{End}(V)$ and $\lambda \in k$, and put $W_\lambda = \mathrm{Im}(\varphi - \lambda)^n$. Then

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

both summands are φ-invariant, and $\varphi - \lambda$ restricts to an isomorphism of W_λ onto itself. In particular, λ is not an eigenvalue of $\varphi|_{W_\lambda}$.

Proof. Apply Lemma 7.6.5 to the operator $\varphi - \lambda$ with $m = n$. This gives $V = \ker(\varphi - \lambda)^n \oplus \operatorname{Im}(\varphi - \lambda)^n = V_\lambda \oplus W_\lambda$ and the statement about $\varphi - \lambda$ on W_λ. Both summands are kernels or images of a polynomial in φ, so they are φ-invariant (Proposition 7.3.3). □

Proposition 7.6.10 (Independence of generalized eigenspaces)

Let $\lambda_1, \dots, \lambda_\ell \in k$ be distinct. If $v_1 + \cdots + v_\ell = 0$ with $v_i \in V_{\lambda_i}$, then $v_i = 0$ for all i. That is, the sum $V_{\lambda_1} + \cdots + V_{\lambda_\ell}$ is direct.

Proof. We show $v_1 = 0$; the same argument applies to every v_i. Suppose $v_1 \neq 0$, and let $r \geqslant 0$ be the largest integer with $w = (\varphi - \lambda_1)^r(v_1) \neq 0$. Then $(\varphi - \lambda_1)(w) = 0$, i.e., $\varphi(w) = \lambda_1 w$. Consider the operator

$$T = (\varphi - \lambda_2)^n \circ \cdots \circ (\varphi - \lambda_\ell)^n \circ (\varphi - \lambda_1)^r.$$

Its factors commute (Proposition 7.2.2). For $j \geqslant 2$, $T(v_j) = 0$, because we may apply $(\varphi - \lambda_j)^n$ first and $(\varphi - \lambda_j)^n(v_j) = 0$. On the other hand, $(\varphi - \lambda_j)(w) = (\lambda_1 - \lambda_j)w$, so

$$T(v_1) = (\varphi - \lambda_2)^n \circ \cdots \circ (\varphi - \lambda_\ell)^n(w) = \prod_{j=2}^{\ell} (\lambda_1 - \lambda_j)^n \, w \neq 0.$$

Applying T to $v_1 + \cdots + v_\ell = 0$ gives $T(v_1) = 0$, a contradiction. Hence $v_1 = 0$. □

7.6.3 The Generalized Eigenspace Decomposition

The following theorem is the main structural result for operators on finite-dimensional vector spaces over algebraically closed fields. Its proof is an induction on the dimension built on the invariant splitting, and it does not use the Cayley–Hamilton Theorem.

Theorem 7.6.11 (Generalized eigenspace decomposition)

Let V be a finite-dimensional vector space over an algebraically closed field k, and let $\varphi \in \operatorname{End}(V)$. Let $\lambda_1, \dots, \lambda_\ell$ be the distinct eigenvalues of φ. Then

$$V = V_{\lambda_1} \oplus V_{\lambda_2} \oplus \cdots \oplus V_{\lambda_\ell}.$$

In particular, $\sum_{i=1}^{\ell} \dim V_{\lambda_i} = \dim V$.

Proof. By Proposition 7.6.10, the sum of the V_{λ_i} is direct, so it remains to show that they span V. We use induction on $\dim V$. If $V = \{0\}$, there is nothing to prove.

Let $\dim V = n \geqslant 1$ and assume the result for all spaces of smaller dimension. By Theorem 7.5.1, φ has at least one eigenvalue, say λ_1. By Proposition 7.6.9,

$$V = V_{\lambda_1} \oplus U, \qquad U = \operatorname{Im}(\varphi - \lambda_1)^n,$$

with U invariant under φ. Since $V_{\lambda_1} \neq \{0\}$, we have $\dim U < \dim V$. By the induction hypothesis applied to $\varphi|_U$,

$$U = U_{\mu_1} \oplus \cdots \oplus U_{\mu_s},$$

where $\mu_1, \ldots, \mu_s$ are the eigenvalues of $\varphi|_U$ and $U_\mu = \ker(\varphi|_U - \mu)^{\dim U}$ are its generalized eigenspaces. An eigenvector of $\varphi|_U$ is an eigenvector of φ, so each μ_j is one of $\lambda_1, \ldots, \lambda_\ell$; and $U_{\mu_j} \subseteq \ker(\varphi - \mu_j)^n = V_{\mu_j}$. Therefore

$$V = V_{\lambda_1} + U_{\mu_1} + \cdots + U_{\mu_s} \subseteq V_{\lambda_1} + V_{\lambda_2} + \cdots + V_{\lambda_\ell},$$

so the generalized eigenspaces span V. The dimension formula follows because the sum is direct. □

Corollary 7.6.12 (The image of $(\varphi - \lambda_i)^n$)
In the situation of Theorem 7.6.11, with $n = \dim V$,

$$\operatorname{Im}(\varphi - \lambda_i)^n = \bigoplus_{j \neq i} V_{\lambda_j} \quad \text{for each } i.$$

Proof. For $j \neq i$, the operator $\varphi - \lambda_i$ maps V_{λ_j} isomorphically onto itself (Proposition 7.6.8(d)), so $V_{\lambda_j} = (\varphi - \lambda_i)^n(V_{\lambda_j}) \subseteq \operatorname{Im}(\varphi - \lambda_i)^n$. Hence $\bigoplus_{j \neq i} V_{\lambda_j} \subseteq \operatorname{Im}(\varphi - \lambda_i)^n$. By Proposition 7.6.9 and Theorem 7.6.11, both sides have dimension $n - \dim V_{\lambda_i}$, so they are equal. □

Combining the decomposition with triangular form gives a matrix in which each eigenvalue occupies its own block.

Corollary 7.6.13 (Block-triangular form)
Let k be algebraically closed, $\varphi \in \operatorname{End}(V)$, and let $\lambda_1, \ldots, \lambda_\ell$ be the distinct eigenvalues of φ, with $d_i = \dim V_{\lambda_i}$. There is a basis of V in which the matrix of φ is block diagonal,

$$M(\varphi) = \begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_\ell \end{pmatrix}, \qquad A_i = \begin{pmatrix} \lambda_i & & * \\ & \ddots & \\ 0 & & \lambda_i \end{pmatrix} \in M_{d_i \times d_i}(k),$$

where each A_i is upper triangular with all diagonal entries equal to λ_i.

Proof. Each V_{λ_i} is a nonzero invariant subspace. By Theorem 7.5.2, it has a basis in which $\varphi|_{V_{\lambda_i}}$ has an upper-triangular matrix A_i. The diagonal entries of A_i are eigenvalues of $\varphi|_{V_{\lambda_i}}$ (Corollary 7.5.4), and the only such eigenvalue is λ_i (Proposition 7.6.8(d)). Concatenating these bases gives a basis of $V = \bigoplus V_{\lambda_i}$, in which the matrix is block diagonal by Theorem 7.3.4. □

Theorem 7.6.14 (Algebraic multiplicity equals $\dim V_\lambda$)

Let k be algebraically closed, $\varphi \in \operatorname{End}(V)$, and let $\lambda_1, \dots, \lambda_\ell$ be the distinct eigenvalues of φ. Then

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

Hence the algebraic multiplicity of each eigenvalue λ_i is $\dim V_{\lambda_i}$, and the algebraic multiplicities add up to $\dim V$.

Proof. The matrix of Corollary 7.6.13 is upper triangular, and λ_i appears on its diagonal exactly $\dim V_{\lambda_i}$ times. Apply Corollary 7.5.4(b). □

Corollary 7.6.15 (Geometric and algebraic multiplicity)

Let k be algebraically closed and λ an eigenvalue of $\varphi \in \operatorname{End}(V)$. Then

$$1 \leqslant \dim E(\lambda, \varphi) \leqslant \dim V_\lambda = (\text{algebraic multiplicity of } \lambda).$$

The operator φ is diagonalizable if and only if the geometric multiplicity equals the algebraic multiplicity for every eigenvalue, i.e., if and only if $E(\lambda, \varphi) = V_\lambda$ for every eigenvalue λ.

Proof. The inequalities follow from $\{0\} \neq E(\lambda, \varphi) \subseteq V_\lambda$ and Theorem 7.6.14. By Theorem 7.4.13, φ is diagonalizable if and only if $\sum_i \dim E(\lambda_i, \varphi) = \dim V$, and $\dim V = \sum_i \dim V_{\lambda_i}$ by Theorem 7.6.11. Since $\dim E(\lambda_i, \varphi) \leqslant \dim V_{\lambda_i}$ for each i, the two sums agree if and only if every term agrees. □

Example 7.6.16 (Decomposition for a 3×3 matrix)

Consider $\varphi : \mathbb{C}^3 \to \mathbb{C}^3$ with matrix

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

The matrix is upper triangular, so $\chi_A(x) = (x-2)^2(x-3)$: the eigenvalue 2 has algebraic multiplicity 2 and the eigenvalue 3 has algebraic multiplicity 1. For $\lambda = 2$:

$$A - 2I = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad (A-2I)^2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and $(A-2I)^3 = (A-2I)^2$. So $V_2 = \ker(A-2I)^3 = \ker(A-2I)^2 = \mathrm{Span}\{e_1, e_2\}$, which has dimension 2, while the eigenspace $E(2,\varphi) = \mathrm{Span}\{e_1\}$ has dimension 1.
For $\lambda = 3$: the powers of $A - 3I$ all have kernel $\mathrm{Span}\{e_3\}$, so $V_3 = E(3,\varphi) = \mathrm{Span}\{e_3\}$, of dimension 1.
Indeed, $\mathbb{C}^3 = V_2 \oplus V_3$, the dimensions agree with the algebraic multiplicities, and φ is not diagonalizable.

7.7 Nilpotent Operators

Nilpotent operators (Definition 7.6.4) are the building blocks for understanding non-diagonalizable maps. On each generalized eigenspace V_λ, the operator $\varphi - \lambda$ is nilpotent, so understanding nilpotent operators is the key to understanding all operators over an algebraically closed field.

7.7.1 Basic Properties

Example 7.7.1 (Basic examples of nilpotent operators)

(a) The zero operator on $V \neq \{0\}$ is nilpotent with index 1.

(b) The shift operator on k^n with matrix $\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}$, which sends $e_1 \mapsto 0$ and $e_j \mapsto e_{j-1}$ for $j \geqslant 2$, is nilpotent with index n.

(c) Any strictly upper-triangular matrix N (zeros on and below the diagonal) is nilpotent: N maps $\mathrm{Span}(e_1, \ldots, e_j)$ into $\mathrm{Span}(e_1, \ldots, e_{j-1})$, so $N^n = 0$.

Proposition 7.7.2 (Characterizations of nilpotent operators)

For $\varphi \in \mathrm{End}(V)$ with $\dim V = n$, the following are equivalent:

(a) φ is nilpotent.

(b) $\varphi^n = 0$.

(c) $V = \mathrm{gker}(\varphi)$.

If k is algebraically closed and $V \neq \{0\}$, they are also equivalent to:

(d) 0 is the only eigenvalue of φ.

Proof. (b) $\Rightarrow$ (a) is clear. (a) $\Rightarrow$ (c): if $\varphi^m = 0$, every vector is killed by φ^m. (c) $\Rightarrow$ (b): $\mathrm{gker}(\varphi) = \ker(\varphi^n)$ by Proposition 7.6.2(b), so $V = \ker(\varphi^n)$.

(a) $\Rightarrow$ (d): If $\varphi^m = 0$ and $\varphi(v) = \lambda v$ with $v \neq 0$, then $0 = \varphi^m(v) = \lambda^m v$, so $\lambda = 0$. And 0 is an eigenvalue: φ is not injective, since otherwise φ^m would be injective on $V \neq \{0\}$.

(d) $\Rightarrow$ (c): By the generalized eigenspace decomposition (Theorem 7.6.11), $V = \bigoplus_\lambda V_\lambda$ over the eigenvalues λ. If the only eigenvalue is 0, then $V = V_0 = \mathrm{gker}(\varphi)$. □

NOTE Over a field that is not algebraically closed, the absence of nonzero eigenvalues does not imply nilpotence: the rotation by 90° of $\mathbb{R}^2$ has no eigenvalues at all. What remains true over every field is that φ is nilpotent if and only if $\chi_\varphi(x) = x^n$. If φ is nilpotent, the next corollary gives a strictly upper-triangular matrix for φ, and then $\chi_\varphi(x) = x^n$ by Corollary 7.5.4(b); the converse is a consequence of the Cayley–Hamilton Theorem (Theorem 7.9.1).

Corollary 7.7.3 (Nilpotent operators have strictly upper-triangular matrices)

Let V be finite-dimensional. If $\varphi \in \mathrm{End}(V)$ is nilpotent, then there exists a basis of V in which the matrix of φ is strictly upper triangular. This holds over every field.

Proof. We prove this by induction on $n = \dim V$. If $n = 0$ there is nothing to prove, and if $\varphi = 0$ any basis works.

Let $\varphi \neq 0$ and $n \geqslant 1$. Consider $W = \mathrm{Im}(\varphi)$. We have $W \subsetneq V$: otherwise φ would be surjective, hence bijective (Corollary 6.1.16), and then no power of φ could be zero on $V \neq \{0\}$. The subspace W is φ-invariant, and $\varphi|_W$ is nilpotent. By induction, W has a basis $(w_1, \ldots, w_r)$ with $\varphi(w_j) \in \mathrm{Span}(w_1, \ldots, w_{j-1})$ for each j. Extend it to a basis $(w_1, \ldots, w_r, u_{r+1}, \ldots, u_n)$ of V. Each $\varphi(u_j)$ lies in $W = \mathrm{Span}(w_1, \ldots, w_r)$, which is spanned by vectors preceding u_j. So in this

basis every basis vector is mapped into the span of the preceding ones, which means the matrix is strictly upper triangular. □

7.7.2 The Nilpotent Basis Construction

The key to understanding nilpotent operators is building a basis adapted to the "chains" of vectors connected by φ.

Definition 7.7.4 (Chains)

Let $\varphi \in \mathrm{End}(V)$ be nilpotent. A ***chain*** (or ***Jordan chain***) of length m is a sequence of vectors $v_1, v_2, \ldots, v_m$ with $v_1 \neq 0$ such that:

$$\varphi(v_m) = v_{m-1}, \quad \varphi(v_{m-1}) = v_{m-2}, \quad \ldots, \quad \varphi(v_2) = v_1, \quad \varphi(v_1) = 0.$$

The vector v_m is called the ***head*** of the chain, and v_1 is the ***tail***. Equivalently, the chain is $\varphi^{m-1}(v_m), \ldots, \varphi(v_m), v_m$ for a vector v_m with $\varphi^{m-1}(v_m) \neq 0$ and $\varphi^m(v_m) = 0$.

NOTE In a chain, $v_1 \in \ker(\varphi)$, $v_2 \in \ker(\varphi^2) \setminus \ker(\varphi)$, and generally $v_j \in \ker(\varphi^j) \setminus \ker(\varphi^{j-1})$, since $\varphi^j(v_j) = \varphi(v_1) = 0$ while $\varphi^{j-1}(v_j) = v_1 \neq 0$. The chain records how φ moves vectors toward the kernel.

Lemma 7.7.5 (Key lemma for chain construction)

Let φ be nilpotent and $j \geqslant 1$. Suppose $U \subseteq \ker(\varphi^{j+1})$ is a subspace satisfying $U \cap \ker(\varphi^j) = \{0\}$. Then:

(a) $\varphi|_U$ is injective.

(b) $\varphi(U) \subseteq \ker(\varphi^j)$.

(c) $\varphi(U) \cap \ker(\varphi^{j-1}) = \{0\}$.

Proof. (a) If $v \in U$ and $\varphi(v) = 0$, then $v \in U \cap \ker(\varphi) \subseteq U \cap \ker(\varphi^j) = \{0\}$, using $j \geqslant 1$.

(b) For $v \in U \subseteq \ker(\varphi^{j+1})$, we have $\varphi^{j+1}(v) = 0$, so $\varphi^j(\varphi(v)) = 0$, meaning $\varphi(v) \in \ker(\varphi^j)$.

(c) Let $v \in U$ with $\varphi(v) \in \ker(\varphi^{j-1})$. Then $\varphi^j(v) = \varphi^{j-1}(\varphi(v)) = 0$, so $v \in U \cap \ker(\varphi^j) = \{0\}$, hence $\varphi(v) = 0$. □

Theorem 7.7.6 (Structure theorem for nilpotent operators)

Let $V \neq \{0\}$ be finite-dimensional and let $\varphi \in \mathrm{End}(V)$ be nilpotent with index m (so $\varphi^m = 0$ but $\varphi^{m-1} \neq 0$). Then there exists a basis of V of the form

$$\{v_{1,1}, \dots, v_{1,m_1};\ v_{2,1}, \dots, v_{2,m_2};\ \dots;\ v_{\ell,1}, \dots, v_{\ell,m_\ell}\}$$

where each sequence $(v_{i,1}, \dots, v_{i,m_i})$ is a chain:

$$\varphi(v_{i,j}) = v_{i,j-1} \text{ for } j > 1, \quad \varphi(v_{i,1}) = 0.$$

The chain lengths satisfy $m = m_1 \geqslant m_2 \geqslant \cdots \geqslant m_\ell \geqslant 1$ and $\sum_{i=1}^{\ell} m_i = \dim V$, and the number of chains is $\ell = \dim \ker(\varphi)$.

Proof. We construct subspaces U_j for $j = m, m-1, \dots, 1$, working down from $\ker(\varphi^m) = V$, together with bases $\mathcal{B}_j$ of U_j, such that

$$\ker(\varphi^j) = \ker(\varphi^{j-1}) \oplus U_j \quad \text{and} \quad \varphi(\mathcal{B}_{j+1}) \subseteq \mathcal{B}_j,$$

with φ injective on $\mathcal{B}_{j+1}$.

First step: Let U_m be a complement of $\ker(\varphi^{m-1})$ in $V = \ker(\varphi^m)$ (extend a basis of $\ker(\varphi^{m-1})$ to a basis of V, and let U_m be the span of the added vectors $\mathcal{B}_m$). Since $\varphi^{m-1} \neq 0$, $U_m \neq \{0\}$. These vectors will be the heads of the chains of length m.

Inductive step: Suppose U_{j+1} and $\mathcal{B}_{j+1}$ have been constructed for some $j \geqslant 1$, with $\ker(\varphi^{j+1}) = \ker(\varphi^j) \oplus U_{j+1}$. By Lemma 7.7.5 applied to $U = U_{j+1}$, the map φ is injective on U_{j+1}, so $\varphi(\mathcal{B}_{j+1})$ is a linearly independent set of $|\mathcal{B}_{j+1}|$ vectors spanning $\varphi(U_{j+1})$; moreover $\varphi(U_{j+1}) \subseteq \ker(\varphi^j)$ and $\varphi(U_{j+1}) \cap \ker(\varphi^{j-1}) = \{0\}$. Hence a basis of $\ker(\varphi^{j-1})$ together with $\varphi(\mathcal{B}_{j+1})$ is linearly independent; extend it by vectors $\mathcal{C}_j$ to a basis of $\ker(\varphi^j)$. Set

$$\mathcal{B}_j = \varphi(\mathcal{B}_{j+1}) \cup \mathcal{C}_j, \qquad U_j = \mathrm{Span}(\mathcal{B}_j).$$

Then $\ker(\varphi^j) = \ker(\varphi^{j-1}) \oplus U_j$. The new vectors $\mathcal{C}_j$ (possibly none) will be the heads of chains of length j.

Assembling the basis: Since $\ker(\varphi^0) = \{0\}$, the decompositions give

$$\begin{aligned} V = \ker(\varphi^m) &= \ker(\varphi^{m-1}) \oplus U_m \\ &= \ker(\varphi^{m-2}) \oplus U_{m-1} \oplus U_m = \cdots = U_1 \oplus U_2 \oplus \cdots \oplus U_m, \end{aligned}$$

so $\mathcal{B}_1 \cup \cdots \cup \mathcal{B}_m$ is a basis of V. Put $\mathcal{C}_m = \mathcal{B}_m$. Each $u \in \mathcal{C}_j$ gives the chain $\varphi^{j-1}(u), \dots, \varphi(u), u$ of length j: its vectors lie in $\mathcal{B}_1, \dots, \mathcal{B}_j$, $\varphi^j(u) = 0$ because $\mathcal{B}_1 \subseteq U_1 = \ker(\varphi)$, and $\varphi^{j-1}(u) \neq 0$ because it is a basis vector. Every element of $\mathcal{B}_i$ either belongs to $\mathcal{C}_i$ or is the image under φ of exactly one element of $\mathcal{B}_{i+1}$;

following these images upward, every basis vector lies in exactly one chain. Thus the basis is partitioned into chains.

The longest chains have length m (their heads form $\mathcal{C}_m = \mathcal{B}_m \neq \varnothing$), and ordering the chains by decreasing length gives $m = m_1 \geqslant \cdots \geqslant m_\ell$. Each chain has exactly one vector in $\mathcal{B}_1$, its tail, so $\ell = |\mathcal{B}_1| = \dim U_1 = \dim \ker(\varphi)$. □

NOTE The chain lengths $(m_1, m_2, \ldots, m_\ell)$ form a partition of $n = \dim V$. We will see in Proposition 7.7.10 that this partition depends only on φ: the choices made in the construction do not affect it.

7.7.3 Jordan Blocks for Nilpotent Operators

The chain basis gives the matrix of a nilpotent operator a very specific form.

Definition 7.7.7 (Nilpotent Jordan block)

A ***nilpotent Jordan block*** of size m is the $m \times m$ matrix

$$J_m(0) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

It has 1's on the superdiagonal and 0's elsewhere.

Corollary 7.7.8 (Matrix form of nilpotent operators)

With respect to a chain basis ordered as $(v_{1,1}, \ldots, v_{1,m_1}, v_{2,1}, \ldots, v_{2,m_2}, \ldots)$, the matrix of a nilpotent operator φ is block diagonal:

$$M(\varphi) = \begin{pmatrix} J_{m_1}(0) & & 0 \\ & J_{m_2}(0) & \\ 0 & & \ddots \end{pmatrix} = J_{m_1}(0) \oplus J_{m_2}(0) \oplus \cdots \oplus J_{m_\ell}(0).$$

Proof. Each chain spans a φ-invariant subspace, and V is the direct sum of these subspaces, so the matrix is block diagonal (Theorem 7.3.4). Within the chain $(v_{i,1}, \ldots, v_{i,m_i})$, we have $\varphi(v_{i,j}) = v_{i,j-1}$ for $j > 1$ and $\varphi(v_{i,1}) = 0$: the j-th column of the block has a single 1, in row $j-1$. This is exactly $J_{m_i}(0)$. □

Example 7.7.9 (Nilpotent operator with partition $(3,2,1)$)

Consider $\varphi : k^6 \to k^6$ nilpotent with chain lengths $3,2,1$. A chain basis might be:

$$(v_{1,1}, v_{1,2}, v_{1,3}, v_{2,1}, v_{2,2}, v_{3,1})$$

where $\varphi(v_{1,3}) = v_{1,2}$, $\varphi(v_{1,2}) = v_{1,1}$, $\varphi(v_{1,1}) = 0$, $\varphi(v_{2,2}) = v_{2,1}$, $\varphi(v_{2,1}) = 0$, and $\varphi(v_{3,1}) = 0$.
The matrix is:

$$\mathcal{M}(\varphi) = \left(\begin{array}{ccc|cc|c} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right) = J_3(0) \oplus J_2(0) \oplus J_1(0).$$

Proposition 7.7.10 (Uniqueness of the chain lengths)

Let V be finite-dimensional, $\varphi \in \mathrm{End}(V)$ nilpotent, and let $d_j = \dim \ker(\varphi^j)$ for $j \geqslant 0$. In any basis of V consisting of chains for φ, the number of chains of length $\geqslant j$ is $d_j - d_{j-1}$, and the number of chains of length exactly j is $(d_j - d_{j-1}) - (d_{j+1} - d_j)$. In particular, the partition $(m_1, \dots, m_\ell)$ of Theorem 7.7.6 is uniquely determined by φ.

Proof. Let $(v_{i,p})$ be a basis consisting of chains of lengths $m_1, \dots, m_\ell$, with $\varphi(v_{i,p}) = v_{i,p-1}$ and $v_{i,0} = 0$. Then $\varphi^j(v_{i,p}) = v_{i,p-j}$ if $p > j$, and $\varphi^j(v_{i,p}) = 0$ if $p \leqslant j$. The vectors $v_{i,p-j}$ with $p > j$ are distinct basis vectors, hence linearly independent. So $\varphi^j\big(\sum c_{i,p} v_{i,p}\big) = \sum_{p>j} c_{i,p} v_{i,p-j}$ vanishes exactly when $c_{i,p} = 0$ for all $p > j$, and

$$\ker(\varphi^j) = \mathrm{Span}\{v_{i,p} : p \leqslant j\}, \qquad d_j = \sum_{i=1}^{\ell} \min(m_i, j).$$

Therefore $d_j - d_{j-1}$ is the number of i with $m_i \geqslant j$. The number of chains of length exactly j is the number of length $\geqslant j$ minus the number of length $\geqslant j+1$. Since the d_j are determined by φ, so is the partition. $\square$

NOTE This dimension-counting argument is the key to proving uniqueness of the Jordan form in general. It also shows that two nilpotent $n \times n$ matrices are similar if and only if they have the same chain lengths: similar matrices have kernels of powers of the same dimensions, and two nilpotent matrices

with the same chain lengths are both similar to the same block matrix $J_{m_1}(0) \oplus \cdots \oplus J_{m_\ell}(0)$.

7.8 Jordan Normal Form

We now combine the generalized eigenspace decomposition with the structure theory of nilpotent operators to obtain the Jordan normal form—the "best possible" matrix representation for an operator over an algebraically closed field.

7.8.1 Jordan Blocks

Definition 7.8.1 (Jordan block)

A ***Jordan block*** of size m with eigenvalue λ is the $m \times m$ matrix

$$J_m(\lambda) = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & \cdots & 0 & \lambda \end{pmatrix} = \lambda I_m + J_m(0).$$

It has λ on the diagonal, 1's on the superdiagonal, and 0's elsewhere.

Example 7.8.2 (Small Jordan blocks)

$$J_1(\lambda) = (\lambda), \quad J_2(\lambda) = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}, \quad J_3(\lambda) = \begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix}.$$

A 1×1 Jordan block is just a scalar, acting on a line spanned by an eigenvector. Every Jordan block, of any size, contains exactly one eigenvector up to scaling, namely its first basis vector; the remaining basis vectors of a block of size $\geqslant 2$ are generalized eigenvectors that are not eigenvectors.

Proposition 7.8.3 (Properties of Jordan blocks)

Let $J = J_m(\lambda)$. Then:

(a) The only eigenvalue of J is λ, with algebraic multiplicity m.

(b) The eigenspace $E(\lambda, J) = \ker(J - \lambda I) = \operatorname{Span}\{e_1\}$ has dimension 1.

(c) The generalized eigenspace for λ is all of k^m.
(d) $J - \lambda I = J_m(0)$ is nilpotent with index m.

Proof. (a) J is upper triangular with all diagonal entries equal to λ, so $\chi_J(x) = (x-\lambda)^m$ (Corollary 7.5.4).

(b) $J - \lambda I = J_m(0)$ sends $e_1 \mapsto 0$ and $e_j \mapsto e_{j-1}$ for $j \geqslant 2$, so its kernel is spanned by e_1.

(c) Since $(J-\lambda I)^m = J_m(0)^m = 0$, every vector is in $\ker(J-\lambda I)^m$.

(d) $J_m(0)^{m-1} \neq 0$ (it maps $e_m \mapsto e_1$) but $J_m(0)^m = 0$. □

NOTE A Jordan block $J_m(\lambda)$ represents the action of an operator on a single chain of generalized eigenvectors. The chain is $(e_1, e_2, \ldots, e_m)$ where $Je_j = \lambda e_j + e_{j-1}$ for $j > 1$ and $Je_1 = \lambda e_1$. Equivalently, $(J-\lambda I)e_j = e_{j-1}$ for $j > 1$ and $(J-\lambda I)e_1 = 0$.

7.8.2 The Jordan Normal Form Theorem

Theorem 7.8.4 (Jordan normal form)

Let V be a finite-dimensional vector space over an algebraically closed field k, and let $\varphi \in \mathrm{End}(V)$. Then there exists a basis of V with respect to which the matrix of φ is block diagonal:

$$\mathcal{M}(\varphi) = \begin{pmatrix} J_{m_1}(\lambda_1) & & 0 \\ & J_{m_2}(\lambda_2) & \\ 0 & & \ddots \end{pmatrix}$$

where each block is a Jordan block. This matrix is called the ***Jordan normal form*** (or ***Jordan canonical form***) of φ. The scalars λ_i that occur are the eigenvalues of φ (possibly repeated).

Proof. By the generalized eigenspace decomposition (Theorem 7.6.11), $V = V_{\lambda_1} \oplus \cdots \oplus V_{\lambda_r}$ where $\lambda_1, \ldots, \lambda_r$ are the distinct eigenvalues. Each V_{λ_i} is φ-invariant, so by Theorem 7.3.4 we can work on each generalized eigenspace separately.

On V_{λ_i}, the operator $\psi_i = \varphi|_{V_{\lambda_i}} - \lambda_i \mathrm{id}$ is nilpotent (Proposition 7.6.8(c)). By the structure theorem for nilpotent operators, there is a chain basis of V_{λ_i} in which ψ_i has matrix $J_{p_1}(0) \oplus J_{p_2}(0) \oplus \cdots$.

Since $\varphi|_{V_{\lambda_i}} = \psi_i + \lambda_i \mathrm{id}$, the matrix of $\varphi|_{V_{\lambda_i}}$ in this basis is

$$(J_{p_1}(0) + \lambda_i I) \oplus (J_{p_2}(0) + \lambda_i I) \oplus \cdots = J_{p_1}(\lambda_i) \oplus J_{p_2}(\lambda_i) \oplus \cdots .$$

Combining the bases for all generalized eigenspaces gives a basis of V in which φ has the desired block diagonal form. □

NOTE The Jordan form groups blocks by eigenvalue: all blocks $J_m(\lambda_i)$ for a given λ_i together correspond to the decomposition of V_{λ_i} into chains. The total size of the blocks with eigenvalue λ_i equals $\dim V_{\lambda_i}$, which is the algebraic multiplicity of λ_i (Theorem 7.6.14), and the number of these blocks is $\dim E(\lambda_i, \varphi)$, the geometric multiplicity. In particular, φ is diagonalizable if and only if all its Jordan blocks have size 1.

7.8.3 Uniqueness of the Jordan Form

Theorem 7.8.5 (Uniqueness of Jordan normal form)

The Jordan normal form of φ is unique up to permutation of the blocks. More precisely, in any basis in which the matrix of φ is a direct sum of Jordan blocks, the number of blocks $J_m(\lambda)$ of size exactly m with eigenvalue λ is

$$(d_m(\lambda) - d_{m-1}(\lambda)) - (d_{m+1}(\lambda) - d_m(\lambda)) = 2d_m(\lambda) - d_{m-1}(\lambda) - d_{m+1}(\lambda),$$

where $d_j(\lambda) = \dim \ker\big((\varphi - \lambda)^j\big)$.

Proof. Fix a basis in which φ is a direct sum of Jordan blocks and fix $\lambda \in k$. Let W_λ be the span of the basis vectors belonging to blocks with eigenvalue λ, and W' the span of the others; both are φ-invariant and $V = W_\lambda \oplus W'$.

We claim $W_\lambda = V_\lambda$. On W_λ, $\varphi - \lambda$ is a direct sum of blocks $J_p(0)$, hence nilpotent, so $W_\lambda \subseteq V_\lambda$. On W', the matrix of $\varphi - \lambda$ is upper triangular with nonzero diagonal entries $\mu - \lambda$, so $\varphi - \lambda$ is invertible on W' (Lemma 7.5.3). If $v = w + w'$ with $w \in W_\lambda, w' \in W'$ lies in V_λ, then $0 = (\varphi-\lambda)^n(v) = (\varphi-\lambda)^n(w')$, forcing $w' = 0$. So $V_\lambda \subseteq W_\lambda$.

Thus the blocks with eigenvalue λ form a chain basis of V_λ for the nilpotent operator $\psi = (\varphi - \lambda)|_{V_\lambda}$, with chain lengths equal to the block sizes. For $j \geqslant 1$, $\ker(\varphi - \lambda)^j \subseteq V_\lambda$, so $\dim\ker(\psi^j) = d_j(\lambda)$; this also holds for $j = 0$. By Proposition 7.7.10, the number of chains of length exactly m is $(d_m(\lambda) - d_{m-1}(\lambda)) - (d_{m+1}(\lambda) - d_m(\lambda))$. These dimensions are intrinsic to φ, so the block structure is uniquely determined. □

Corollary 7.8.6 (Similarity criterion)

Let k be algebraically closed and $A, B \in \mathcal{M}_{n\times n}(k)$. The following are equivalent:

(a) A and B are similar.

(b) A and B have the same Jordan normal form, up to the order of the blocks.

(c) For each $\lambda \in k$ and each $j \geqslant 1$, $\dim\ker\big((A-\lambda I)^j\big) = \dim\ker\big((B-\lambda I)^j\big)$.

(The Jordan normal form of a matrix A means that of the operator L_A on k^n.)

Proof. (a) $\Rightarrow$ (c): If $B = P^{-1}AP$, then $(B-\lambda I)^j = P^{-1}(A-\lambda I)^j P$, and $x \mapsto P^{-1}x$ maps $\ker\big((A-\lambda I)^j\big)$ isomorphically onto $\ker\big((B-\lambda I)^j\big)$.

(c) $\Rightarrow$ (b): By Theorem 7.8.5, the numbers $\dim\ker\big((A-\lambda I)^j\big)$ determine the Jordan form.

(b) $\Rightarrow$ (a): A matrix is similar to its Jordan normal form J (Corollary 5.3.4), and permuting the blocks of J amounts to reordering the basis, which gives a similar matrix. Similarity is an equivalence relation, so A and B are similar. □

NOTE This gives a complete set of similarity invariants over an algebraically closed field. Trace and determinant alone do not distinguish all similarity classes ($\left(\begin{smallmatrix}2&1\\0&2\end{smallmatrix}\right)$ and $2I$ share both), but the full sequence of kernel dimensions does.

7.8.4 Computing the Jordan Form

Given a matrix A over an algebraically closed field, here is the procedure to find its Jordan form:

Step 1: Find eigenvalues. Compute $\chi_A(x) = \det(xI - A)$ and factor it as $\prod_i (x-\lambda_i)^{n_i}$. The λ_i are the eigenvalues, with algebraic multiplicities n_i; the blocks with eigenvalue λ_i have total size n_i.

Step 2: For each eigenvalue λ, determine block sizes. Compute $d_j = \dim\ker(A-\lambda I)^j$ for $j = 1, 2, \ldots$ until $d_j = n_i$ (after which the sequence is constant). The number of blocks of size $\geqslant j$ is $d_j - d_{j-1}$. In particular:

- Number of blocks = d_1 = geometric multiplicity.
- Number of blocks of size $\geqslant 2$ is $d_2 - d_1$.
- Number of blocks of size exactly j is $(d_j - d_{j-1}) - (d_{j+1} - d_j)$.

Step 3: Assemble the Jordan form. List all Jordan blocks grouped by eigenvalue.

Example 7.8.7 (Computing Jordan form)

Let $A = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}$.

Step 1: $\chi_A(x) = (x-2)^3$, so $\lambda = 2$ with multiplicity 3.

Step 2:

$$A - 2I = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad (A-2I)^2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad (A-2I)^3 = 0.$$

So $d_1 = 1$, $d_2 = 2$, $d_3 = 3$.
Number of blocks = $d_1 = 1$. Size of the single block = 3.
Step 3: The Jordan form is $J_3(2)$. (The matrix is already in Jordan form.)

Example 7.8.8 (Two blocks with the same eigenvalue)

Let $A = \begin{pmatrix} 3 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 3 \end{pmatrix}$.

Step 1: $\chi_A(x) = (x-3)^4$, so $\lambda = 3$ with multiplicity 4.

Step 2:

$$A - 3I = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

$\ker(A-3I) = \text{Span}\{e_1, e_3\}$, so $d_1 = 2$. $(A-3I)^2 = 0$, so $d_2 = 4$.
Number of blocks = 2. Blocks of size $\geqslant 2$: $d_2 - d_1 = 2$. So both blocks have size 2.
Step 3: The Jordan form is $J_2(3) \oplus J_2(3)$.

Example 7.8.9 (Multiple eigenvalues)

Let $A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}$.

Step 1: $\chi_A(x) = (x-1)^2(x-2)^2$. Eigenvalues: $\lambda_1 = 1$, $\lambda_2 = 2$, each with multiplicity 2.
Step 2 for $\lambda = 1$:

$$A - I = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad (A-I)^2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

So $\ker(A - I) = \operatorname{Span}\{e_1\}$ and $\ker(A-I)^2 = \operatorname{Span}\{e_1, e_2\}$: $d_1 = 1$, $d_2 = 2$. One block of size 2: $J_2(1)$.
Step 2 for $\lambda = 2$:

$$A - 2I = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

So $\ker(A - 2I) = \operatorname{Span}\{e_3, e_4\}$ and $d_1 = 2$, already the multiplicity. Two blocks of size 1: $J_1(2) \oplus J_1(2)$.

Step 3: The Jordan form is $J_2(1) \oplus J_1(2) \oplus J_1(2) = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}$.

NOTE In the last example, $\lambda = 2$ contributes two 1×1 blocks because its eigenspace has dimension 2 (geometric = algebraic multiplicity). In contrast, $\lambda = 1$ has a 2×2 block because its geometric multiplicity (1) is less than its algebraic multiplicity (2).

7.9 The Cayley–Hamilton Theorem and the Minimal Polynomial

The characteristic polynomial (Definition 7.4.7) encodes the eigenvalues and their algebraic multiplicities. In this section we show that every operator satisfies its own characteristic polynomial, and we study the "smallest" polynomial relation satisfied by an operator.

7.9.1 The Cayley–Hamilton Theorem

Theorem 7.9.1 (Cayley–Hamilton)

Let V be a finite-dimensional vector space over any field k, and let $\varphi \in \operatorname{End}(V)$ with characteristic polynomial $\chi_\varphi(x)$. Then

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

That is, substituting φ for x in $\chi_\varphi(x)$ gives the zero operator. Equivalently, $\chi_A(A) = 0$ for every square matrix A.

Proof. We give two proofs. The first works over an algebraically closed field and explains the theorem; the second works over every field.

Proof 1 (k algebraically closed): Let $\lambda_1, \dots, \lambda_\ell$ be the distinct eigenvalues of φ and $n_i = \dim V_{\lambda_i}$. By Theorem 7.6.14, $\chi_\varphi(x) = \prod_{i=1}^{\ell} (x - \lambda_i)^{n_i}$, so

$$\chi_\varphi(\varphi) = (\varphi - \lambda_1)^{n_1} \circ \cdots \circ (\varphi - \lambda_\ell)^{n_\ell},$$

and the factors commute (Proposition 7.2.2). Let $v \in V_{\lambda_i}$. The restriction of $\varphi - \lambda_i$ to V_{λ_i} is a nilpotent operator on a space of dimension n_i, so its n_i-th power is zero (Proposition 7.7.2). Applying the factor $(\varphi - \lambda_i)^{n_i}$ first, we get $\chi_\varphi(\varphi)(v) = 0$. Since $V = \bigoplus_i V_{\lambda_i}$ (Theorem 7.6.11), $\chi_\varphi(\varphi) = 0$.

Proof 2 (any field, using the adjugate): Let $A = \mathcal{M}(\varphi, \mathcal{B})$ for a basis $\mathcal{B}$. Sending an operator to its matrix in $\mathcal{B}$ is a ring isomorphism $\operatorname{End}(V) \to \mathcal{M}_{n\times n}(k)$ (Corollary 5.2.10), and it sends $p(\varphi)$ to $p(A)$; so it suffices to show $\chi_A(A) = 0$.

Let $M = xI - A \in \mathcal{M}_{n\times n}(k[x])$, and let $\operatorname{adj}(M)$ be its adjugate. The adjugate identity (Theorem 5.4.16), which holds for matrices over the commutative ring $k[x]$ (Example 5.4.25), gives

$$M \cdot \operatorname{adj}(M) = \det(M)\, I = \chi_A(x)\, I.$$

Each entry of $\operatorname{adj}(M)$ is $\pm$ the determinant of an $(n-1)\times(n-1)$ submatrix of M, which by the Leibniz formula is a polynomial in x of degree at most $n-1$. So we can write

$$\operatorname{adj}(M) = B_{n-1}x^{n-1} + \cdots + B_1 x + B_0, \qquad B_j \in \mathcal{M}_{n\times n}(k).$$

Write also $\chi_A(x) = c_n x^n + \cdots + c_1 x + c_0$ (with $c_n = 1$), and set $B_{-1} = B_n = 0$. Expanding,

$$(xI - A)\sum_{j=0}^{n-1} B_j x^j = \sum_{j=0}^{n} (B_{j-1} - AB_j)\, x^j.$$

Both sides of the adjugate identity are matrices of polynomials, so their coefficients agree:

$$B_{j-1} - AB_j = c_j I \qquad (0 \leqslant j \leqslant n).$$

Multiply the j-th equation on the left by A^j and add:

$$\chi_A(A) = \sum_{j=0}^{n} c_j A^j = \sum_{j=0}^{n} \left(A^j B_{j-1} - A^{j+1} B_j\right).$$

The sum telescopes: the term $A^{j+1}B_j$ subtracted at index j is added back at index $j+1$, so all that remains is $A^0 B_{-1} - A^{n+1} B_n = 0$. Hence $\chi_A(A) = 0$. □

NOTE The Cayley–Hamilton Theorem does *not* follow from "plugging in A for x in $\det(xI - A)$." That substitution produces the scalar $\det(A - A) = 0$, while $\chi_A(A)$ is a matrix; the two have nothing to do with each other a priori. The theorem requires proof.

Proof 2 is not the only route to arbitrary fields. Every field k is contained in an algebraically closed field (a theorem we do not prove here), and the polynomial χ_A and the matrix $\chi_A(A)$ are computed the same way whether the entries of A are regarded in k or in the larger field. So Proof 1, applied over the larger field, also gives $\chi_A(A) = 0$.

Corollary 7.9.2 (Expression for the inverse)

If $\varphi \in \mathrm{End}(V)$ is invertible, then φ^{-1} can be expressed as a polynomial in φ.

Proof. Write $\chi_\varphi(x) = x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0$. Then $c_0 = (-1)^n \det(\varphi) \neq 0$ (Proposition 7.4.8(a) and Theorem 5.4.8).

By Cayley–Hamilton: $\varphi^n + c_{n-1}\varphi^{n-1} + \cdots + c_1\varphi + c_0 \mathrm{id} = 0$.

Rearranging: $\varphi \circ (\varphi^{n-1} + c_{n-1}\varphi^{n-2} + \cdots + c_1 \mathrm{id}) = -c_0 \mathrm{id}$.

Thus $\varphi^{-1} = -\frac{1}{c_0}(\varphi^{n-1} + c_{n-1}\varphi^{n-2} + \cdots + c_1 \mathrm{id})$. □

7.9.2 The Minimal Polynomial

Definition 7.9.3 (Minimal polynomial)

The ***minimal polynomial*** $\mu_\varphi(x)$ of $\varphi \in \mathrm{End}(V)$, V finite-dimensional, is the monic polynomial of smallest degree such that $\mu_\varphi(\varphi) = 0$. (It exists and is unique by the next proposition.) For a square matrix A, μ_A is defined in the same way.

Proposition 7.9.4 (Existence, uniqueness, and divisibility)

Let $\varphi \in \mathrm{End}(V)$ with V finite-dimensional.

(a) There exists a nonzero polynomial $p(x)$ with $p(\varphi) = 0$.

(b) Among all such polynomials, there is a unique monic one of minimal degree.

(c) $\mu_\varphi(x)$ divides every polynomial $p(x)$ satisfying $p(\varphi) = 0$.

Proof. (a) Cayley–Hamilton gives $\chi_\varphi(\varphi) = 0$, and χ_φ is monic, hence nonzero.

(b) Let d be the minimal degree of a nonzero polynomial annihilating φ; dividing by the leading coefficient, there is a monic one. If $p(x)$ and $q(x)$ are both monic of degree d with $p(\varphi) = q(\varphi) = 0$, then $(p-q)(\varphi) = 0$ and $\deg(p-q) < d$. By minimality, $p - q = 0$.

(c) Given $p(x)$ with $p(\varphi) = 0$, divide by μ_φ with remainder: $p(x) = q(x)\mu_\varphi(x) + r(x)$ where $r = 0$ or $\deg r < \deg \mu_\varphi$. (Since μ_φ is monic of degree d, if p has degree $e \geqslant d$ and leading coefficient a, then $p - ax^{e-d}\mu_\varphi$ has smaller degree; repeating this produces q and r.) Then $r(\varphi) = p(\varphi) - q(\varphi)\mu_\varphi(\varphi) = 0 - 0 = 0$. By minimality of μ_φ, we must have $r = 0$, so $\mu_\varphi \mid p$. □

Theorem 7.9.5 (Characteristic and minimal polynomials)

Let $\varphi \in \mathrm{End}(V)$, V finite-dimensional, over any field k.

(a) $\mu_\varphi(x)$ divides $\chi_\varphi(x)$.

(b) For $\lambda \in k$: $\mu_\varphi(\lambda) = 0 \iff \chi_\varphi(\lambda) = 0 \iff \lambda$ is an eigenvalue of φ. So μ_φ and χ_φ have the same roots in k.

Proof. (a) By Cayley–Hamilton, $\chi_\varphi(\varphi) = 0$. By Proposition 7.9.4(c), $\mu_\varphi \mid \chi_\varphi$.

(b) The second equivalence is Proposition 7.4.8(d). If λ is an eigenvalue, there exists $v \neq 0$ with $\varphi(v) = \lambda v$. For any polynomial p, $p(\varphi)(v) = p(\lambda)v$. Taking $p = \mu_\varphi$ gives $0 = \mu_\varphi(\lambda)v$, so $\mu_\varphi(\lambda) = 0$. Conversely, if $\mu_\varphi(\lambda) = 0$, then $\chi_\varphi(\lambda) = 0$ because μ_φ divides χ_φ. □

Over an algebraically closed field, the minimal polynomial can be read off from the Jordan form.

Proposition 7.9.6 (Minimal polynomial and Jordan block sizes)

Let k be algebraically closed, $\varphi \in \mathrm{End}(V)$, and let $\lambda_1, \ldots, \lambda_\ell$ be the distinct eigenvalues of φ. For each i, let m_i be the index of nilpotency of $(\varphi - \lambda_i)|_{V_{\lambda_i}}$,

i.e., the smallest m with $V_{\lambda_i} = \ker(\varphi - \lambda_i)^m$. Then

$$\mu_\varphi(x) = \prod_{i=1}^{\ell} (x - \lambda_i)^{m_i}, \qquad 1 \leqslant m_i \leqslant \dim V_{\lambda_i},$$

and m_i is the size of the largest Jordan block with eigenvalue λ_i.

Proof. Let $p(x) = \prod_i (x - \lambda_i)^{m_i}$. On V_{λ_i}, the factor $(\varphi - \lambda_i)^{m_i}$ is zero; since the factors of $p(\varphi)$ commute, $p(\varphi)$ vanishes on each V_{λ_i}, hence on $V = \bigoplus_i V_{\lambda_i}$. Thus $\mu_\varphi \mid p$.

Now fix i and write $\mu_\varphi(x) = (x - \lambda_i)^e h(x)$ with $h(\lambda_i) \neq 0$ (divide by $x - \lambda_i$ as long as possible, Lemma 4.3.14). Since k is algebraically closed, h is a constant times a product of factors $x - c$ with $c \neq \lambda_i$ (Proposition 4.3.17), and each $\varphi - c$ is invertible on the invariant subspace V_{λ_i} (Proposition 7.6.8(d)); so $h(\varphi)$ is invertible on V_{λ_i}. From $0 = \mu_\varphi(\varphi) = h(\varphi) \circ (\varphi - \lambda_i)^e$ we get $(\varphi - \lambda_i)^e = 0$ on V_{λ_i}, so $e \geqslant m_i$. Factoring μ_φ into linear factors and using Lemma 7.4.11, $x - \lambda_i$ occurs at least m_i times for each i, so $\deg \mu_\varphi \geqslant \sum_i m_i = \deg p$. As μ_φ divides p and both are monic, $\mu_\varphi = p$.

We have $m_i \geqslant 1$ because $V_{\lambda_i} \neq \{0\}$, and $m_i \leqslant \dim V_{\lambda_i}$ by Proposition 7.7.2. Finally, in a Jordan basis the blocks with eigenvalue λ_i form a chain basis of V_{λ_i} for $\psi = (\varphi - \lambda_i)|_{V_{\lambda_i}}$ (proof of Theorem 7.8.5), and ψ^j kills a chain of length p if and only if $j \geqslant p$. So the index of nilpotency of ψ is the length of the longest chain, i.e., the size of the largest block. □

Example 7.9.7 (Minimal polynomial examples)

(a) For $A = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} = 2I$: $\chi_A(x) = (x-2)^2$, but $\mu_A(x) = x - 2$ since $A - 2I = 0$.

(b) For $A = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$: $\chi_A(x) = (x-2)^2$. Since $A - 2I = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \neq 0$, we need $\mu_A(x) = (x-2)^2 = \chi_A(x)$.

(c) For $A = \operatorname{diag}(1,1,2,2,2)$: $\chi_A(x) = (x-1)^2(x-2)^3$, but $\mu_A(x) = (x-1)(x-2)$ since $(A - I)(A - 2I) = 0$ and, by Theorem 7.9.5(b), both 1 and 2 must be roots of μ_A.

7.9.3 Diagonalizability and the Minimal Polynomial

The minimal polynomial detects diagonalizability over every field. The proof uses a dimension inequality.

Lemma 7.9.8 (Kernel of a product)
Let $S, T \in \mathrm{End}(V)$, V finite-dimensional. Then $\dim\ker(S\circ T) \leqslant \dim\ker(S) + \dim\ker(T)$. Consequently, $\dim\ker(T_1 \circ \cdots \circ T_m) \leqslant \sum_i \dim\ker(T_i)$.

Proof. Let $K = \ker(S \circ T)$. The map T sends K into $\ker(S)$, and the kernel of $T|_K$ is $\ker(T) \cap K = \ker(T)$. By Rank–Nullity for $T|_K$,

$$\dim K = \dim\ker(T) + \dim T(K) \leqslant \dim\ker(T) + \dim\ker(S).$$

The second statement follows by induction on m. □

Theorem 7.9.9 (Diagonalizability via the minimal polynomial)
Let $\varphi \in \mathrm{End}(V)$, V finite-dimensional, over any field k. Then φ is diagonalizable if and only if $\mu_\varphi(x)$ is a product of distinct linear factors:

$$\mu_\varphi(x) = (x - \lambda_1)(x - \lambda_2)\cdots(x - \lambda_m)$$

with $\lambda_1, \ldots, \lambda_m \in k$ distinct.

Proof. ($\Rightarrow$) Let φ be diagonalizable with distinct eigenvalues $\lambda_1, \ldots, \lambda_m$, and let $p(x) = \prod_{i=1}^{m}(x - \lambda_i)$. Then $p(\varphi) = 0$: on $E(\lambda_j, \varphi)$ the factor $\varphi - \lambda_j$ vanishes, the factors commute, and these eigenspaces span V. Hence $\mu_\varphi \mid p$. On the other hand, every λ_i is a root of μ_φ by Theorem 7.9.5(b). Write $\mu_\varphi = (x - \lambda_1)g_1$ (Lemma 4.3.14); for $i \geqslant 2$, $0 = \mu_\varphi(\lambda_i) = (\lambda_i - \lambda_1)g_1(\lambda_i)$ gives $g_1(\lambda_i) = 0$, and repeating, p divides μ_φ. Two monic polynomials that divide each other are equal, so $\mu_\varphi = p$.

($\Leftarrow$) Suppose $\mu_\varphi = \prod_{i=1}^{m}(x - \lambda_i)$ with distinct λ_i. By Theorem 7.9.5(b), the λ_i are exactly the eigenvalues of φ. Since $\mu_\varphi(\varphi) = (\varphi - \lambda_1) \circ \cdots \circ (\varphi - \lambda_m) = 0$, Lemma 7.9.8 gives

$$\dim V = \dim\ker \mu_\varphi(\varphi) \leqslant \sum_{i=1}^{m} \dim\ker(\varphi - \lambda_i) = \sum_{i=1}^{m} \dim E(\lambda_i, \varphi).$$

The reverse inequality holds by Corollary 7.4.6, so $\sum_i \dim E(\lambda_i, \varphi) = \dim V$, and φ is diagonalizable by Theorem 7.4.13. □

NOTE Over an algebraically closed field this can also be read off from the Jordan form: by Proposition 7.9.6, μ_φ has distinct linear factors if and only if every $m_i = 1$, i.e., if and only if all Jordan blocks have size 1.

7.9.4 Exercises

Question 7.9.10 (Exercises on Jordan form and minimal polynomials)

(a) Let $\varphi : \mathbb{C}^4 \to \mathbb{C}^4$ have characteristic polynomial $\chi_\varphi(x) = (x-1)^2(x-2)^2$. List all possible Jordan normal forms for φ.

(b) Prove that if A and B are similar matrices, then $\mu_A = \mu_B$.

(c) Find the Jordan normal form and minimal polynomial of

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

(d) Let $\varphi \in \mathrm{End}(V)$ satisfy $\varphi^2 = \varphi$, with V finite-dimensional over any field. Prove that φ is diagonalizable and find its possible eigenvalues.

(e) Suppose $\varphi \in \mathrm{End}(V)$ has minimal polynomial $\mu_\varphi(x) = (x-1)(x-2)(x-3)$. What can you conclude about φ?

Solution. **(a)** The eigenvalues are $\lambda_1 = 1$ and $\lambda_2 = 2$, each with algebraic multiplicity 2, so for each eigenvalue the Jordan blocks have sizes summing to 2. Possibilities for each:

- One block of size 2: $J_2(\lambda)$
- Two blocks of size 1: $J_1(\lambda) \oplus J_1(\lambda)$

Combining independently for each eigenvalue, the possible Jordan forms (up to the order of the blocks) are:

(i) $J_2(1) \oplus J_2(2)$
(ii) $J_2(1) \oplus J_1(2) \oplus J_1(2)$
(iii) $J_1(1) \oplus J_1(1) \oplus J_2(2)$
(iv) $J_1(1) \oplus J_1(1) \oplus J_1(2) \oplus J_1(2)$ (diagonal matrix)

(b) If $B = P^{-1}AP$, then $B^j = P^{-1}A^jP$ for all j, so for any polynomial p:

$$p(B) = p(P^{-1}AP) = P^{-1}p(A)P.$$

So $p(B) = 0$ iff $p(A) = 0$. The two matrices are annihilated by the same polynomials, so the monic annihilating polynomial of least degree is the same for both.

(c) The matrix is already in Jordan form: $A = J_3(0)$.
Characteristic polynomial: $\chi_A(x) = x^3$.

Minimal polynomial: We need the smallest j with $A^j = 0$. Compute $A^2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \neq 0$ and $A^3 = 0$. So $\mu_A(x) = x^3 = \chi_A(x)$, in agreement with Proposition 7.9.6 (one block, of size 3).

(d) The condition $\varphi^2 = \varphi$ means $\varphi \circ (\varphi - \mathrm{id}) = 0$. Thus $p(x) = x(x-1)$ annihilates φ, and μ_φ divides $x(x-1)$.

A monic divisor of $x(x-1)$ of degree 1 has the form $x - c$ with c a root of $x(x-1)$, so the monic divisors are 1, x, $x-1$, and $x(x-1)$. Each is a product of distinct linear factors. By Theorem 7.9.9, φ is diagonalizable.

The eigenvalues are roots of μ_φ (Theorem 7.9.5), which divides $x(x-1)$. So the eigenvalues are among $\{0, 1\}$.

(Such operators are called ***projections*** or ***idempotents***. They decompose $V = \ker(\varphi) \oplus \mathrm{Im}(\varphi)$ and project onto $\mathrm{Im}(\varphi)$ along $\ker(\varphi)$; compare the exercises of Section 6.1.)

(e) Since $\mu_\varphi(x) = (x-1)(x-2)(x-3)$ is a product of three distinct linear factors:

- φ is diagonalizable (Theorem 7.9.9).
- The eigenvalues are exactly $1, 2, 3$ (the roots of μ_φ, by Theorem 7.9.5).
- $\dim V \geqslant 3$ (at least one eigenvector for each eigenvalue, and these are linearly independent).
- $V = E(1, \varphi) \oplus E(2, \varphi) \oplus E(3, \varphi)$.

The eigenspace dimensions can be any positive integers summing to $\dim V$; each is at least 1.

7.10 Real Vector Spaces and Complexification

Let V be a real vector space of dimension n and $\varphi \in \mathrm{End}(V)$. Since $\mathbb{R}$ is not algebraically closed, φ might have no eigenvalues, and then it cannot be put in triangular or Jordan form. Nevertheless, every real operator has an invariant subspace of dimension 1 or 2. The approach is to work over $\mathbb{C}$, which is algebraically closed, by extending the scalars of V.

7.10.1 The Complexification of a Real Vector Space

Definition 7.10.1 (Complexification)

Let V be a vector space over $\mathbb{R}$. The ***complexification*** of V is the set $V_{\mathbb{C}} = V \times V$, whose elements (v, w) we write as $v + iw$ ($v, w \in V$), with addition

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

and multiplication by complex scalars

$$(a + ib)(v + iw) = (av - bw) + i(bv + aw), \qquad a, b \in \mathbb{R}.$$

We identify $v \in V$ with $v + i0 \in V_{\mathbb{C}}$. For $\mathbf{v} = v + iw$, the vectors $v = \mathrm{Re}(\mathbf{v})$ and $w = \mathrm{Im}(\mathbf{v})$ are its ***real*** and ***imaginary parts***.

The scalar multiplication is what one gets by expanding $(a + ib)(v + iw)$ formally with $i^2 = -1$. Every element of $V_{\mathbb{C}}$ is uniquely $v + iw$ with $v, w \in V$, so two elements are equal if and only if their real parts and their imaginary parts are equal. Note that $V \subseteq V_{\mathbb{C}}$ is closed under addition and real scalars, but not under complex scalars: $i \cdot v = 0 + iv$.

Proposition 7.10.2 ($V_{\mathbb{C}}$ is a complex vector space)

With these operations, $V_{\mathbb{C}}$ is a vector space over $\mathbb{C}$. If $(e_1, \dots, e_n)$ is a basis of V over $\mathbb{R}$, then $(e_1, \dots, e_n)$ (that is, $(e_1 + i0, \dots, e_n + i0)$) is a basis of $V_{\mathbb{C}}$ over $\mathbb{C}$. In particular, $\dim_{\mathbb{C}} V_{\mathbb{C}} = \dim_{\mathbb{R}} V$.

Proof. Under addition, $V_{\mathbb{C}} = V \times V$ is an abelian group. The scalar axioms are direct computations; for example, associativity: for $\alpha = a + ib$ and $\beta = c + id$,

$$\begin{aligned} \alpha\big(\beta(v + iw)\big) &= \alpha\big((cv - dw) + i(dv + cw)\big) \\ &= \big((ac - bd)v - (ad + bc)w\big) + i\big((ad + bc)v + (ac - bd)w\big) \\ &= (\alpha\beta)(v + iw), \end{aligned}$$

since $\alpha\beta = (ac - bd) + i(ad + bc)$. The distributive laws and $1 \cdot \mathbf{v} = \mathbf{v}$ are checked in the same way.

Spanning: if $v = \sum_j a_j e_j$ and $w = \sum_j b_j e_j$ with $a_j, b_j \in \mathbb{R}$, then $\sum_j (a_j + ib_j)e_j = \sum_j a_j e_j + i\sum_j b_j e_j = v + iw$.

Independence: if $\sum_j (a_j + ib_j)e_j = 0$, then comparing real and imaginary parts, $\sum_j a_j e_j = 0$ and $\sum_j b_j e_j = 0$, so all $a_j = b_j = 0$. $\square$

Example 7.10.3

The complexification of $\mathbb{R}^n$ is $\mathbb{C}^n$: the element $x+iy$ of $(\mathbb{R}^n)_{\mathbb{C}}$, with $x, y \in \mathbb{R}^n$, corresponds to the vector $(x_1 + iy_1, \dots, x_n + iy_n) \in \mathbb{C}^n$, and the operations match.

Definition 7.10.4 (Complex conjugation)

The ***complex conjugate*** of $\mathbf{v} = v + iw \in V_{\mathbb{C}}$ is $\bar{\mathbf{v}} = v - iw$.

Conjugation is additive, satisfies $\overline{\bar{\mathbf{v}}} = \mathbf{v}$, and is *conjugate-linear*: $\overline{\alpha\mathbf{v}} = \bar{\alpha}\,\bar{\mathbf{v}}$ for $\alpha \in \mathbb{C}$. Indeed, for $\alpha = a + ib$, both $\overline{\alpha(v + iw)}$ and $(a - ib)(v - iw)$ equal $(av - bw) - i(bv + aw)$. A vector $\mathbf{v} \in V_{\mathbb{C}}$ lies in V if and only if $\bar{\mathbf{v}} = \mathbf{v}$, and $\operatorname{Re}(\mathbf{v}) = \frac{1}{2}(\mathbf{v} + \bar{\mathbf{v}})$, $\operatorname{Im}(\mathbf{v}) = \frac{1}{2i}(\mathbf{v} - \bar{\mathbf{v}})$.

7.10.2 Complexifying Linear Maps

Proposition 7.10.5 (The complexification of a linear map)

Let $\varphi : V \to W$ be an $\mathbb{R}$-linear map between real vector spaces, and define

$$\varphi_{\mathbb{C}} : V_{\mathbb{C}} \to W_{\mathbb{C}}, \qquad \varphi_{\mathbb{C}}(v + iw) = \varphi(v) + i\varphi(w).$$

Then:

(a) $\varphi_{\mathbb{C}}$ is $\mathbb{C}$-linear, and it agrees with φ on $V \subseteq V_{\mathbb{C}}$.

(b) $\varphi_{\mathbb{C}}$ commutes with conjugation: $\varphi_{\mathbb{C}}(\bar{\mathbf{v}}) = \overline{\varphi_{\mathbb{C}}(\mathbf{v})}$.

(c) $(\mathrm{id}_V)_{\mathbb{C}} = \mathrm{id}_{V_{\mathbb{C}}}$, and $(\psi \circ \varphi)_{\mathbb{C}} = \psi_{\mathbb{C}} \circ \varphi_{\mathbb{C}}$ for every $\mathbb{R}$-linear $\psi : W \to U$.

(d) If V and W are finite-dimensional with bases $\mathcal{B}$ and $\mathcal{C}$, then $\mathcal{M}(\varphi_{\mathbb{C}}, \mathcal{B}, \mathcal{C}) = \mathcal{M}(\varphi, \mathcal{B}, \mathcal{C})$, where $\mathcal{B}$ and $\mathcal{C}$ are regarded as complex bases of $V_{\mathbb{C}}$ and $W_{\mathbb{C}}$ (Proposition 7.10.2). In particular, for an operator $\varphi \in \operatorname{End}(V)$, $\chi_{\varphi_{\mathbb{C}}} = \chi_{\varphi}$, a polynomial with real coefficients.

Proof. (a) Additivity is clear. For $a, b \in \mathbb{R}$,

$$\begin{aligned}\varphi_{\mathbb{C}}\big((a + ib)(v + iw)\big) &= \varphi_{\mathbb{C}}\big((av - bw) + i(bv + aw)\big)\\ &= \big(a\varphi(v) - b\varphi(w)\big) + i\big(b\varphi(v) + a\varphi(w)\big),\end{aligned}$$

which is $(a + ib)\big(\varphi(v) + i\varphi(w)\big)$. And $\varphi_{\mathbb{C}}(v + i0) = \varphi(v) + i0$.

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

(c) Clearly $(\mathrm{id}_V)_{\mathbb{C}}(v + iw) = v + iw$, and $(\psi \circ \varphi)_{\mathbb{C}}(v + iw) = \psi(\varphi(v)) + i\psi(\varphi(w)) = \psi_{\mathbb{C}}\big(\varphi(v) + i\varphi(w)\big) = \psi_{\mathbb{C}}(\varphi_{\mathbb{C}}(v + iw))$.

(d) If $\mathcal{B} = (e_1, \dots, e_n)$ and $\mathcal{C} = (f_1, \dots, f_m)$, then $\varphi_{\mathbb{C}}(e_j) = \varphi(e_j) = \sum_i a_{ij} f_i$, so the j-th columns of the two matrices agree. The characteristic polynomial of an operator is computed from its matrix. □

NOTE Part (c) says that $V \mapsto V_{\mathbb{C}}$, $\varphi \mapsto \varphi_{\mathbb{C}}$ is a functor from real to complex vector spaces, in the language of Chapter 8 (Example 8.2.4).

Now let $\varphi \in \mathrm{End}(V)$ be an operator on a finite-dimensional real space V. Then $\varphi_{\mathbb{C}}$ is an operator on a complex vector space, so it is guaranteed to have an eigenvector if $V \neq \{0\}$ (and it has generalized eigenspaces and a Jordan form). Its eigenvectors tell us about φ.

Proposition 7.10.6 (Eigenvectors of the complexification)

Let $\varphi \in \mathrm{End}(V)$, where V is a real vector space, and let $\mathbf{v} = v + iw$ be an eigenvector of $\varphi_{\mathbb{C}}$ with eigenvalue $\lambda \in \mathbb{C}$.

(a) If $\lambda \in \mathbb{R}$, then v and w lie in $E(\lambda, \varphi)$, and at least one of them is nonzero; so λ is an eigenvalue of φ. Conversely, every eigenvalue of φ is an eigenvalue of $\varphi_{\mathbb{C}}$. Thus the real eigenvalues of $\varphi_{\mathbb{C}}$ are exactly the eigenvalues of φ.

(b) If $\lambda = a + ib$ with $b \neq 0$, then $\bar{\mathbf{v}}$ is an eigenvector of $\varphi_{\mathbb{C}}$ with eigenvalue $\bar{\lambda}$; the vectors $v, w \in V$ are linearly independent; and $U = \mathrm{Span}_{\mathbb{R}}(v, w)$ is a two-dimensional φ-invariant subspace of V, with

$$\varphi(v) = av - bw, \quad \varphi(w) = bv + aw, \qquad \mathcal{M}\big(\varphi|_U, (v, w)\big) = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}.$$

Proof. In both cases, $\varphi_{\mathbb{C}}(\mathbf{v}) = \varphi(v) + i\varphi(w)$ and $\lambda\mathbf{v} = (a + ib)(v + iw) = (av - bw) + i(bv + aw)$, where $\lambda = a + ib$. Comparing real and imaginary parts,

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

(a) If $b = 0$, this says $\varphi(v) = av$ and $\varphi(w) = aw$, and v, w are not both zero because $\mathbf{v} \neq 0$. Conversely, an eigenvector $v \in V$ of φ is an eigenvector of $\varphi_{\mathbb{C}}$ with the same eigenvalue, by Proposition 7.10.5(a).

(b) By Proposition 7.10.5(b), $\varphi_{\mathbb{C}}(\bar{\mathbf{v}}) = \overline{\lambda\mathbf{v}} = \bar{\lambda}\bar{\mathbf{v}}$, and $\bar{\mathbf{v}} \neq 0$. Since $\lambda \neq \bar{\lambda}$, the eigenvectors $\mathbf{v}$ and $\bar{\mathbf{v}}$ are linearly independent over $\mathbb{C}$ (Theorem 7.4.4). As $v = \frac{1}{2}(\mathbf{v} + \bar{\mathbf{v}})$ and $w = \frac{1}{2i}(\mathbf{v} - \bar{\mathbf{v}})$, and conversely $\mathbf{v} = v + iw$, $\bar{\mathbf{v}} = v - iw$, the complex spans of (v, w) and of $(\mathbf{v}, \bar{\mathbf{v}})$ coincide and are two-dimensional; so v, w are linearly independent over $\mathbb{C}$, and in particular over $\mathbb{R}$. The displayed formulas show $\varphi(U) \subseteq U$, and the columns of the matrix are the coordinates of $\varphi(v) = a \cdot v + (-b) \cdot w$ and $\varphi(w) = b \cdot v + a \cdot w$. □

Theorem 7.10.7 (Invariant subspaces of real operators)

Every operator on a nonzero finite-dimensional real vector space has an invariant subspace of dimension 1 or 2.

Proof. Let $\varphi \in \mathrm{End}(V)$ with $V \neq \{0\}$ finite-dimensional over $\mathbb{R}$. Then $V_{\mathbb{C}}$ is a nonzero finite-dimensional complex vector space, and $\mathbb{C}$ is algebraically closed (the Fundamental Theorem of Algebra, Chapter 4), so $\varphi_{\mathbb{C}}$ has an eigenvector $\mathbf{v}$ (Theorem 7.5.1). If its eigenvalue is real, Proposition 7.10.6(a) gives an eigenvector of φ, which spans a one-dimensional invariant subspace. Otherwise, Proposition 7.10.6(b) gives a two-dimensional invariant subspace. □

Example 7.10.8 (The rotation, revisited)

Let $R = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ act on $\mathbb{R}^2$. It has no eigenvalues, but its complexification (the same matrix acting on $\mathbb{C}^2$) has the eigenvector $\mathbf{v} = (1, -i)^t$ with eigenvalue $\lambda = i$: indeed $R\mathbf{v} = (i, 1)^t = i\,(1, -i)^t$. Here $v = \mathrm{Re}(\mathbf{v}) = (1, 0)^t$, $w = \mathrm{Im}(\mathbf{v}) = (0, -1)^t$, $a = 0$ and $b = 1$. The formulas of Proposition 7.10.6 say $Rv = -w = (0, 1)^t$ and $Rw = v = (1, 0)^t$, which is correct, and $U = \mathbb{R}^2$ with matrix $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ in the basis (v, w). The conjugate $\bar{\mathbf{v}} = (1, i)^t$ is the eigenvector for $\bar{\lambda} = -i$.

NOTE Since $\chi_{\varphi_{\mathbb{C}}} = \chi_{\varphi}$ has real coefficients, its non-real roots come in conjugate pairs, matching the pairing $\mathbf{v} \leftrightarrow \bar{\mathbf{v}}$ of eigenvectors. A real eigenvalue of $\varphi_{\mathbb{C}}$ contributes eigenvectors of φ itself, while each pair $\lambda, \bar{\lambda}$ of non-real eigenvalues contributes a two-dimensional invariant subspace on which φ acts by a matrix $\left(\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix}\right)$, a rotation combined with a scaling. Continuing by induction with the induced operator on V/U (Proposition 7.3.5), as in the proof of Theorem 7.5.2, one can put every real operator in block upper-triangular form with diagonal blocks of size 1 or 2. Complexification returns in Chapter 9, for operators on real inner product spaces, and in Chapter 16, for real representations.

CHAPTER 8

Categories: An Interlude

Many otherwise disparate fields of mathematics share fundamental structures. The language of categories systematizes these common patterns. While category theory won't answer questions specific to a particular area, it suggests useful ways to think about problems by analogy with other settings. We've already encountered categorical ideas implicitly: the universal property of quotient spaces, the behavior of dual maps, and the isomorphism theorems all fit naturally into this framework. In this chapter we set up the basic vocabulary: categories, functors between them, natural transformations between functors, and universal properties.

8.1 The Language of Categories

8.1.1 Objects and Morphisms

The basic data of a category consists of objects and arrows between them, subject to axioms that capture the essential features of function composition.

Definition 8.1.1 (Category)

A ***category*** $\mathcal{C}$ consists of:

(1) A collection of ***objects***, denoted $\mathrm{Ob}(\mathcal{C})$.

(2) For every pair of objects $A, B \in \mathrm{Ob}(\mathcal{C})$, a collection of ***morphisms*** from A to B, denoted $\mathrm{Mor}(A, B)$ or $\mathrm{Hom}_{\mathcal{C}}(A, B)$.

(3) A ***composition rule***: for every triple $A, B, C \in \mathrm{Ob}(\mathcal{C})$, a map

$$\mathrm{Mor}(A, B) \times \mathrm{Mor}(B, C) \longrightarrow \mathrm{Mor}(A, C)$$

sending (f, g) to $g \circ f$.

These must satisfy two axioms:

(i) **Associativity:** For any $f \in \mathrm{Mor}(A, B)$, $g \in \mathrm{Mor}(B, C)$, $h \in \mathrm{Mor}(C, D)$,

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

(ii) **Identity:** For every object A, there exists $\mathrm{id}_A \in \mathrm{Mor}(A, A)$ such that for any $f \in \mathrm{Mor}(A, B)$,

$$f \circ \mathrm{id}_A = f = \mathrm{id}_B \circ f.$$

NOTE The identity morphism id_A is unique: if id'_A also satisfies the identity axiom, then $\mathrm{id}'_A = \mathrm{id}'_A \circ \mathrm{id}_A = \mathrm{id}_A$.

We write $f : A \to B$ to indicate $f \in \mathrm{Mor}(A, B)$, and call A the ***domain*** (or source) and B the ***codomain*** (or target) of f. When several categories are in play we write $\mathrm{Mor}_{\mathcal{C}}(A, B)$ to record the category.

Example 8.1.2 (The category of sets)

The category **Set** has:

- Objects: all sets.
- Morphisms: $\mathrm{Mor}(A, B) = \{f : A \to B \mid f \text{ is a function}\}$.
- Composition: ordinary function composition.

Variants include:

- **FinSet**: objects are finite sets, morphisms are functions.
- $\mathbf{Set}_*$: ***pointed sets***. Objects are pairs (A, a) with $a \in A$ a distinguished element. Morphisms $(A, a) \to (B, b)$ are functions $f : A \to B$ with $f(a) = b$.

Example 8.1.3 (The category of groups)

The category **Grp** has:

- Objects: all groups.
- Morphisms: group homomorphisms.
- Composition: composition of homomorphisms.

The identity axiom holds because $\mathrm{id}_G : G \to G$ is a homomorphism, and a composition of homomorphisms is a homomorphism. Variants:

- **Ab**: abelian groups and group homomorphisms.

- **FinGrp**: finite groups and homomorphisms.

Similarly, we have categories **Ring** (rings and ring homomorphisms) and **CRing** (commutative rings and ring homomorphisms).

Example 8.1.4 (The category of vector spaces)

Fix a field k. The category $\mathbf{Vect}_k$ has:

- Objects: vector spaces over k.
- Morphisms: k-linear maps.
- Composition: composition of linear maps.

The category $\mathbf{FinVect}_k$ has the finite-dimensional vector spaces over k as objects, again with linear maps as morphisms. We've been studying $\mathbf{Vect}_k$ all along: $\mathrm{Hom}(V, W)$ is exactly $\mathrm{Mor}(V, W)$ in this category.

Example 8.1.5 (The category of topological spaces)

The category **Top** has:

- Objects: topological spaces.
- Morphisms: continuous maps.
- Composition: composition of continuous maps.

The pointed variant $\mathbf{Top}_*$ has objects (X, x_0) (a space with a basepoint) and morphisms the continuous maps preserving basepoints. This category is fundamental in algebraic topology.

NOTE In all these examples, objects are "decorated sets"—sets equipped with additional structure (a group operation, a topology, a vector space structure)—and morphisms are functions that respect this structure. This is typical of categories arising in practice, though the abstract definition permits more exotic examples, such as the following one, in which morphisms are not functions at all.

Example 8.1.6 (Posets as categories)

Let $(P, \leqslant)$ be a partially ordered set. Define a category $\mathcal{C}_P$ by:

- $\mathrm{Ob}(\mathcal{C}_P) = P$.

- $\mathrm{Mor}(a,b) = \begin{cases} \{*\} & \text{if } a \leqslant b, \\ \varnothing & \text{otherwise.} \end{cases}$
- Composition: if $a \leqslant b$ and $b \leqslant c$, the unique element of $\mathrm{Mor}(a,b) \times \mathrm{Mor}(b,c)$ is sent to the unique element of $\mathrm{Mor}(a,c)$, which is nonempty by transitivity.

The identity axiom uses reflexivity ($a \leqslant a$), and associativity is automatic since each $\mathrm{Mor}(a,b)$ has at most one element.

8.1.2 Isomorphisms

The notion of "sameness" in a category is captured by isomorphisms.

Definition 8.1.7 (Isomorphism)

Let $\mathcal{C}$ be a category and $A, B \in \mathrm{Ob}(\mathcal{C})$. A morphism $f \in \mathrm{Mor}(A,B)$ is an ***isomorphism*** if there exists $g \in \mathrm{Mor}(B,A)$ such that

$$g \circ f = \mathrm{id}_A \quad \text{and} \quad f \circ g = \mathrm{id}_B.$$

We write $A \cong B$ if such an isomorphism exists, and say A and B are ***isomorphic***.

Proposition 8.1.8 (Uniqueness of inverses)

If $f : A \to B$ is an isomorphism, then the morphism g in the definition is unique. We denote it f^{-1} and call it the ***inverse*** of f.

Proof. Suppose $g, g' \in \mathrm{Mor}(B,A)$ both satisfy the conditions. Then

$$g = g \circ \mathrm{id}_B = g \circ (f \circ g') = (g \circ f) \circ g' = \mathrm{id}_A \circ g' = g'. \qquad \square$$

Proposition 8.1.9 (Properties of isomorphisms)

Let $\mathcal{C}$ be a category.

(i) For any object A, id_A is an isomorphism with $\mathrm{id}_A^{-1} = \mathrm{id}_A$.

(ii) If $f : A \to B$ is an isomorphism, then $f^{-1} : B \to A$ is an isomorphism with $(f^{-1})^{-1} = f$.

(iii) If $f : A \to B$ and $g : B \to C$ are isomorphisms, then $g \circ f : A \to C$ is an isomorphism with $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.

Proof. (i) $\mathrm{id}_A \circ \mathrm{id}_A = \mathrm{id}_A$.

(ii) By definition, $f^{-1} \circ f = \mathrm{id}_A$ and $f \circ f^{-1} = \mathrm{id}_B$. This says exactly that f is an inverse of f^{-1}.

(iii) Compute:

$$(f^{-1} \circ g^{-1}) \circ (g \circ f) = f^{-1} \circ (g^{-1} \circ g) \circ f = f^{-1} \circ \mathrm{id}_B \circ f = f^{-1} \circ f = \mathrm{id}_A.$$

Similarly, $(g \circ f) \circ (f^{-1} \circ g^{-1}) = \mathrm{id}_C$. □

Example 8.1.10 (Isomorphisms in familiar categories)

- In **Set**: isomorphisms are bijections.
- In **Grp**: isomorphisms are group isomorphisms (bijective homomorphisms; the inverse of a bijective homomorphism is automatically a homomorphism).
- In $\mathbf{Vect}_k$: isomorphisms are invertible linear maps.
- In **Top**: isomorphisms are homeomorphisms (continuous bijections with continuous inverse).

In $\mathbf{FinVect}_k$, two objects are isomorphic iff they have the same dimension. In **FinSet**, two objects are isomorphic iff they have the same cardinality.

NOTE In **Top**, a bijective continuous map need not be an isomorphism: the inverse may fail to be continuous. For example, $\mathrm{id} : (\mathbb{R}, \text{discrete}) \to (\mathbb{R}, \text{standard})$ is a continuous bijection but not a homeomorphism. So "isomorphism" in a category can mean more than "bijection on underlying sets."

8.1.3 Automorphism Groups

The self-isomorphisms of an object form a group, connecting category theory back to group theory.

Definition 8.1.11 (Automorphism group)

For an object A in a category $\mathcal{C}$, the ***automorphism group*** of A is

$$\mathrm{Aut}(A) = \{f \in \mathrm{Mor}(A, A) : f \text{ is an isomorphism}\}.$$

Theorem 8.1.12 (Automorphisms form a group)
$\mathrm{Aut}(A)$, equipped with composition, is a group.

Proof.
- **Closure:** If $f, g \in \mathrm{Aut}(A)$, then $g \circ f$ is an isomorphism by Proposition 8.1.9(iii).
- **Associativity:** Inherited from the category axiom.
- **Identity:** $\mathrm{id}_A \in \mathrm{Aut}(A)$ by Proposition 8.1.9(i), and it is an identity element for composition.
- **Inverses:** If $f \in \mathrm{Aut}(A)$, then $f^{-1} \in \mathrm{Aut}(A)$ by Proposition 8.1.9(ii). □

Example 8.1.13 (Automorphism groups in practice)

- In **FinSet**: if $|A| = n$, then $\mathrm{Aut}(A)$ is the group of permutations of A, isomorphic to the symmetric group S_n.
- In $\mathbf{FinVect}_k$: if $\dim V = n$, then $\mathrm{Aut}(V)$ is the group of invertible linear operators on V, isomorphic to $\mathrm{GL}_n(k)$ by choosing a basis.
- In **Grp**: $\mathrm{Aut}(G)$ is the group of group automorphisms of G. For $G = \mathbb{Z}/n$, we have $\mathrm{Aut}(\mathbb{Z}/n) \cong (\mathbb{Z}/n)^\times$: an endomorphism of $\mathbb{Z}/n$ is multiplication by some $\bar{a}$ (it is determined by the image $\bar{a}$ of $\bar{1}$), and it is bijective iff $\bar{a}$ generates $\mathbb{Z}/n$, i.e., iff $\bar{a}$ is a unit.

Exercise 8.1.14 (Isomorphic objects have isomorphic automorphism groups)
Let $f : A \to B$ be an isomorphism in a category $\mathcal{C}$. Show that the map

$$c_f : \mathrm{Aut}(A) \to \mathrm{Aut}(B), \quad g \mapsto f \circ g \circ f^{-1}$$

is a group isomorphism.

Solution. **Well-defined:** If $g \in \mathrm{Aut}(A)$, then $f \circ g \circ f^{-1}$ is a composition of isomorphisms, hence an isomorphism $B \to B$.
Homomorphism: For $g, h \in \mathrm{Aut}(A)$:

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

Bijective: The map $c_{f^{-1}} : \mathrm{Aut}(B) \to \mathrm{Aut}(A)$ is an inverse:

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

Similarly, $c_f(c_{f^{-1}}(h)) = h$.

NOTE This exercise shows that the automorphism group is an "isomorphism invariant": isomorphic objects have isomorphic automorphism groups. The converse is false: non-isomorphic objects can have isomorphic automorphism groups. For instance, in **Grp**, the cyclic groups $\mathbb{Z}$ and $\mathbb{Z}/3$ both have automorphism group $\mathbb{Z}/2$ (generated by inversion $n \mapsto -n$ in the first case and by $\bar{1} \mapsto \bar{2}$ in the second), but $\mathbb{Z} \not\cong \mathbb{Z}/3$.

A category in which every morphism is an isomorphism has a special name.

Definition 8.1.15 (Groupoid)

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

NOTE A groupoid generalizes the notion of a group: a group is the same thing as a groupoid with exactly one object (the group being the automorphism group of that object). In a groupoid, morphisms can be composed only when the target of one matches the source of the other, but when composition is defined, it is associative, has identities, and has inverses. The fundamental groupoid of a topological space, whose objects are the points of the space and whose morphisms are homotopy classes of paths, is a key example in algebraic topology.

Example 8.1.16 (Groupoids)

- The category whose objects are finite sets and whose morphisms are only the bijections (not all functions) is a groupoid.
- Any group G defines a groupoid $\mathbf{B}G$ with one object $*$, with $\mathrm{Mor}(*,*) = G$ and composition given by the group law.
- The category whose objects are groups and whose morphisms are only the group isomorphisms is a groupoid.

8.2 Functors and Natural Transformations

Constructions that turn objects of one kind into objects of another kind, and do so compatibly with morphisms, are formalized by functors.

8.2.1 Functors

Definition 8.2.1 (Functor)

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

- to each object $X \in \mathrm{Ob}(\mathcal{C})$, of an object $F(X) \in \mathrm{Ob}(\mathcal{D})$, and
- to each morphism $f \in \mathrm{Mor}_{\mathcal{C}}(X, Y)$, of a morphism $F(f) \in \mathrm{Mor}_{\mathcal{D}}(F(X), F(Y))$,

such that

(i) $F(\mathrm{id}_X) = \mathrm{id}_{F(X)}$ for every object X, and

(ii) $F(g \circ f) = F(g) \circ F(f)$ for all $f \in \mathrm{Mor}_{\mathcal{C}}(X, Y)$ and $g \in \mathrm{Mor}_{\mathcal{C}}(Y, Z)$.

Example 8.2.2 (Identity and forgetful functors)

- Every category $\mathcal{C}$ has an ***identity functor*** $\mathrm{id}_{\mathcal{C}}$, with $\mathrm{id}_{\mathcal{C}}(X) = X$ and $\mathrm{id}_{\mathcal{C}}(f) = f$.
- The ***forgetful functors*** **Grp** $\to$ **Set**, **Top** $\to$ **Set**, $\mathbf{Vect}_k \to$ **Set**, ... send a group, a topological space, a vector space to its underlying set, and a homomorphism, continuous map, or linear map to the same map regarded as a function. Both axioms hold trivially. Similarly, there are forgetful functors **Ab** $\to$ **Grp** (regard an abelian group as a group) and **Ring** $\to$ **Ab** (keep the additive group of a ring and forget its multiplication).

Example 8.2.3 (The functor $\mathrm{Hom}(V, \cdot)$)

Fix a vector space V over k. For a vector space W, set $F(W) = \mathrm{Hom}(V, W)$. For a linear map $f : W \to W'$, let

$$F(f) : \mathrm{Hom}(V, W) \to \mathrm{Hom}(V, W'), \qquad a \mapsto f \circ a.$$

This map is linear, since $f \circ (a + \lambda b) = f \circ a + \lambda f \circ b$. The axioms hold: $F(\mathrm{id}_W)(a) = \mathrm{id}_W \circ a = a$, and for $g : W' \to W''$,

$$F(g \circ f)(a) = (g \circ f) \circ a = g \circ (f \circ a) = F(g)(F(f)(a)).$$

So $\mathrm{Hom}(V, \cdot)$ is a functor $\mathbf{Vect}_k \to \mathbf{Vect}_k$.

Example 8.2.4 (Complexification)

Complexification (Section 7.10) is a functor $\mathbf{Vect}_{\mathbb{R}} \to \mathbf{Vect}_{\mathbb{C}}$. On objects it sends V to $V_{\mathbb{C}}$. On morphisms it sends an $\mathbb{R}$-linear map $\varphi : V \to W$ to the $\mathbb{C}$-linear map $\varphi_{\mathbb{C}} : V_{\mathbb{C}} \to W_{\mathbb{C}}$, $\varphi_{\mathbb{C}}(v+iw) = \varphi(v)+i\varphi(w)$, of Proposition 7.10.5. The functor axioms are part (c) of that proposition: clearly $(\mathrm{id}_V)_{\mathbb{C}} = \mathrm{id}_{V_{\mathbb{C}}}$, and for $\psi : W \to U$,

$$\begin{aligned}(\psi \circ \varphi)_{\mathbb{C}}(v + iw) &= \psi(\varphi(v)) + i\psi(\varphi(w)) \\ &= \psi_{\mathbb{C}}(\varphi(v) + i\varphi(w)) = (\psi_{\mathbb{C}} \circ \varphi_{\mathbb{C}})(v + iw).\end{aligned}$$

Some constructions reverse the direction of morphisms.

Definition 8.2.5 (Contravariant functor)

A ***contravariant functor*** $F : \mathcal{C} \to \mathcal{D}$ assigns to each object X of $\mathcal{C}$ an object $F(X)$ of $\mathcal{D}$, and to each morphism $f \in \mathrm{Mor}_{\mathcal{C}}(X, Y)$ a morphism $F(f) \in \mathrm{Mor}_{\mathcal{D}}(F(Y), F(X))$ *in the opposite direction*, such that

$$F(\mathrm{id}_X) = \mathrm{id}_{F(X)} \qquad \text{and} \qquad F(g \circ f) = F(f) \circ F(g).$$

Example 8.2.6 (Duality)

The dual space construction of Chapter 6 is a contravariant functor $\mathbf{Vect}_k \to \mathbf{Vect}_k$: it sends V to $V^* = \mathrm{Hom}(V, k)$ and a linear map $\varphi : V \to W$ to its dual map $\varphi^t : W^* \to V^*$, $\varphi^t(\ell) = \ell \circ \varphi$ (Definition 6.2.9). Indeed $\mathrm{id}_V^t(\ell) = \ell$, and for $\psi : W \to U$ and $\ell \in U^*$,

$$(\psi \circ \varphi)^t(\ell) = \ell \circ \psi \circ \varphi = \varphi^t(\ell \circ \psi) = (\varphi^t \circ \psi^t)(\ell).$$

More generally, for a fixed vector space T, the same computation shows that $V \mapsto \mathrm{Hom}(V, T)$, $\varphi \mapsto (a \mapsto a \circ \varphi)$, is a contravariant functor; duality is the case $T = k$.

Functors can be composed: if $F : \mathcal{C} \to \mathcal{D}$ and $G : \mathcal{D} \to \mathcal{E}$ are functors, then $G \circ F$, defined by $(G \circ F)(X) = G(F(X))$ and $(G \circ F)(f) = G(F(f))$, satisfies both axioms. If F and G are both contravariant, the two reversals cancel and $G \circ F$ is covariant: for $f : X \to Y$ and $g : Y \to Z$,

$$G(F(g \circ f)) = G(F(f) \circ F(g)) = G(F(g)) \circ G(F(f)).$$

If exactly one of them is contravariant, $G \circ F$ is contravariant.

Example 8.2.7 (The double dual)

Applying duality twice gives a covariant functor $\mathbf{Vect}_k \to \mathbf{Vect}_k$ sending V to V^{**} and $\varphi : V \to W$ to $\varphi^{tt} = (\varphi^t)^t : V^{**} \to W^{**}$. Explicitly, $\varphi^{tt}(\xi) = \xi \circ \varphi^t$ for $\xi \in V^{**}$.

Proposition 8.2.8 (Functors preserve isomorphisms)

Let $F : \mathcal{C} \to \mathcal{D}$ be a functor (covariant or contravariant). If $f : X \to Y$ is an isomorphism in $\mathcal{C}$, then $F(f)$ is an isomorphism in $\mathcal{D}$, with inverse $F(f^{-1})$. In particular, $X \cong Y$ implies $F(X) \cong F(Y)$.

Proof. In the covariant case, $F(f^{-1}) \circ F(f) = F(f^{-1} \circ f) = F(\mathrm{id}_X) = \mathrm{id}_{F(X)}$, and similarly $F(f) \circ F(f^{-1}) = F(f \circ f^{-1}) = \mathrm{id}_{F(Y)}$. In the contravariant case, $F(f) \circ F(f^{-1}) = F(f^{-1} \circ f) = \mathrm{id}_{F(X)}$ and $F(f^{-1}) \circ F(f) = F(f \circ f^{-1}) = \mathrm{id}_{F(Y)}$. □

NOTE Proposition 8.2.8 is a source of invariants: to show that two objects are *not* isomorphic, it suffices to find a functor that sends them to non-isomorphic objects. The forgetful functor to **Set** already shows, for instance, that groups of different orders are not isomorphic.

8.2.2 Natural Transformations

There is one more layer: morphisms between functors.

Definition 8.2.9 (Natural transformation)

Let $F, G : \mathcal{C} \to \mathcal{D}$ be two (covariant) functors. A ***natural transformation*** t from F to G, written $t : F \Rightarrow G$, is the data, for every object X of $\mathcal{C}$, of a morphism $t_X \in \mathrm{Mor}_{\mathcal{D}}(F(X), G(X))$, such that for all objects X, Y and every $f \in \mathrm{Mor}_{\mathcal{C}}(X, Y)$,

$$t_Y \circ F(f) = G(f) \circ t_X,$$

that is, the square

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

commutes in $\mathcal{D}$. The morphism t_X is the ***component*** of t at X. If every component t_X is an isomorphism, t is a ***natural isomorphism***. Natural

transformations between two contravariant functors are defined in the same way, with the vertical arrows $F(f) : F(Y) \to F(X)$ and $G(f) : G(Y) \to G(X)$ reversed.

The word "natural" has been used informally throughout linear algebra for maps that are defined without choices. The definition makes this precise: a natural map must be compatible with every morphism. The motivating example is the map from a vector space to its double dual.

Theorem 8.2.10 (The double dual is natural)

For each vector space V over k, let $\mathrm{ev}_V : V \to V^{**}$ be the evaluation map, $\mathrm{ev}_V(v)(\ell) = \ell(v)$. Then the maps ev_V are the components of a natural transformation

$$\mathrm{ev} : \mathrm{id}_{\mathbf{Vect}_k} \Rightarrow (\cdot)^{**}$$

from the identity functor to the double dual functor. On $\mathbf{FinVect}_k$ it is a natural isomorphism.

Proof. Let $\varphi : V \to W$ be linear. We must show that the square

$$\begin{array}{ccc} V & \xrightarrow{\mathrm{ev}_V} & V^{**} \\ {\scriptstyle\varphi}\downarrow & & \downarrow{\scriptstyle\varphi^{tt}} \\ W & \xrightarrow[\mathrm{ev}_W]{} & W^{**} \end{array}$$

commutes, i.e., that $\varphi^{tt}(\mathrm{ev}_V(v)) = \mathrm{ev}_W(\varphi(v))$ for all $v \in V$. Both sides are elements of W^{**}, so we evaluate them on an arbitrary $\ell \in W^*$:

$$\varphi^{tt}(\mathrm{ev}_V(v))(\ell) = \mathrm{ev}_V(v)(\varphi^t(\ell)) = \varphi^t(\ell)(v) = \ell(\varphi(v)) = \mathrm{ev}_W(\varphi(v))(\ell).$$

So ev is a natural transformation. If V is finite-dimensional, ev_V is an isomorphism (Theorem 6.2.19), so on $\mathbf{FinVect}_k$ every component is an isomorphism. □

NOTE By contrast, a finite-dimensional space V is also isomorphic to V^*, but every such isomorphism we have seen depends on a choice of basis (or of extra structure, such as an inner product in Chapter 9). The definition explains why no choice-free version can exist in this sense: $V \mapsto V^*$ is contravariant while the identity functor is covariant, so a family of maps $V \to V^*$ cannot even be compared with morphisms in the way the naturality square requires.

Example 8.2.11 (A natural transformation for Hom)
Fix a linear map $\alpha : V' \to V$. For each W, define

$$t_W : \mathrm{Hom}(V, W) \to \mathrm{Hom}(V', W), \qquad a \mapsto a \circ \alpha.$$

Then t is a natural transformation from the functor $\mathrm{Hom}(V, \cdot)$ to the functor $\mathrm{Hom}(V', \cdot)$ of Example 8.2.3: for $f : W \to W'$ and $a \in \mathrm{Hom}(V, W)$, both ways around the square give $f \circ a \circ \alpha$, by associativity of composition.

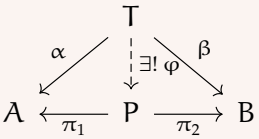
8.3 Universal Properties

Many constructions in mathematics are characterized not by explicit formulas but by *universal properties*—conditions that specify the construction uniquely up to isomorphism. We've already seen this with quotient spaces: the quotient V/U is characterized by the fact that any linear map that kills U factors uniquely through it. Products and coproducts provide two more fundamental examples.

8.3.1 Products and Coproducts

Definition 8.3.1 (Product)

Let $\mathcal{C}$ be a category and $A, B \in \mathrm{Ob}(\mathcal{C})$. A ***product*** of A and B is an object P together with morphisms $\pi_1 : P \to A$ and $\pi_2 : P \to B$ (called ***projections***) satisfying the following ***universal property***:
For any object T and morphisms $\alpha : T \to A$, $\beta : T \to B$, there exists a *unique* morphism $\varphi : T \to P$ such that $\pi_1 \circ \varphi = \alpha$ and $\pi_2 \circ \varphi = \beta$.



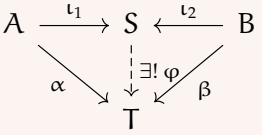
We write $P = A \times B$ and $\varphi = (\alpha, \beta)$ for the unique induced map.

NOTE The universal property says: to give a map into the product is the same as giving a map into each factor. This is the categorical abstraction of the familiar fact that a function $f : T \to A \times B$ is determined by its components $f_1 = \pi_1 \circ f$ and $f_2 = \pi_2 \circ f$.

The dual notion reverses all arrows.

Definition 8.3.2 (Coproduct)

Let $\mathcal{C}$ be a category and $A, B \in \mathrm{Ob}(\mathcal{C})$. A ***coproduct*** (or ***sum***) of A and B is an object S together with morphisms $\iota_1 : A \to S$ and $\iota_2 : B \to S$ (called ***inclusions*** or ***coprojections***) satisfying:
For any object T and morphisms $\alpha : A \to T$, $\beta : B \to T$, there exists a *unique* morphism $\varphi : S \to T$ such that $\varphi \circ \iota_1 = \alpha$ and $\varphi \circ \iota_2 = \beta$.



We write $S = A \amalg B$ or $A + B$ and $\varphi = [\alpha, \beta]$ for the unique induced map.

NOTE The universal property of the coproduct says: to give a map out of the coproduct is the same as giving a map out of each summand. This duality between products and coproducts—reversing all arrows—is a general phenomenon in category theory.

Theorem 8.3.3 (Uniqueness of products and coproducts)

If a product (resp. coproduct) of A and B exists, it is unique up to unique isomorphism. That is, if (P, π_1, π_2) and (P', π_1', π_2') are both products of A and B, then there is a unique isomorphism $\psi : P \to P'$ with $\pi_1' \circ \psi = \pi_1$ and $\pi_2' \circ \psi = \pi_2$; and similarly for coproducts.

Proof. By the universal property of P', applied to the maps $\pi_1 : P \to A$ and $\pi_2 : P \to B$, there exists a unique $\psi : P \to P'$ with $\pi_1' \circ \psi = \pi_1$ and $\pi_2' \circ \psi = \pi_2$.

By the universal property of P, applied to $\pi_1' : P' \to A$ and $\pi_2' : P' \to B$, there exists a unique $\psi' : P' \to P$ with $\pi_1 \circ \psi' = \pi_1'$ and $\pi_2 \circ \psi' = \pi_2'$.

Now $\psi' \circ \psi : P \to P$ satisfies $\pi_1 \circ (\psi' \circ \psi) = \pi_1$ and $\pi_2 \circ (\psi' \circ \psi) = \pi_2$. But id_P also satisfies these equations. By uniqueness in the universal property of P (applied to $T = P$, $\alpha = \pi_1$, $\beta = \pi_2$), we have $\psi' \circ \psi = \mathrm{id}_P$. Similarly, $\psi \circ \psi' = \mathrm{id}_{P'}$. So ψ is an isomorphism, and it is the only morphism with $\pi_1' \circ \psi = \pi_1$ and $\pi_2' \circ \psi = \pi_2$.

The proof for coproducts is the same with all arrows reversed. □

NOTE This proof pattern—using universal properties to construct inverse maps—is standard in category theory. The key is that universal properties give *uniqueness*, which forces the composites to be identities.

Products and coproducts need not exist in every category. When working in a new category, one of the first questions is whether these constructions exist and what form they take.

8.3.2 Products and Coproducts in Familiar Categories

Example 8.3.4 (Products and coproducts in **Set**)

In the category of sets:

- The **product** $A \times B$ is the Cartesian product $\{(a, b) : a \in A, b \in B\}$ with projections $\pi_1(a, b) = a$ and $\pi_2(a, b) = b$.
- The **coproduct** $A \amalg B$ is the disjoint union $A \sqcup B = \{(a, 1) : a \in A\} \cup \{(b, 2) : b \in B\}$ with inclusions $\iota_1(a) = (a, 1)$ and $\iota_2(b) = (b, 2)$.

For the coproduct, given $\alpha : A \to T$ and $\beta : B \to T$, the only possible φ is $\varphi(a, 1) = \alpha(a)$, $\varphi(b, 2) = \beta(b)$. The product case is Exercise 8.3.7(1).

Example 8.3.5 (Products and coproducts in **Grp**)

In the category of groups:

- The **product** $G \times H$ is the direct product, with componentwise operations and the two projections.
- The **coproduct** $G * H$ is the ***free product***, a more complicated construction whose elements are reduced words alternating between nonidentity elements of G and of H. Exercise 8.3.7(2) treats the case $G = H = \mathbb{Z}$, where it is the free group on two generators (Section 3.6).

In the subcategory **Ab** of abelian groups, both the product and the coproduct of G and H are the direct sum $G \oplus H = G \times H$ (the same group, with different universal properties: projections for the product, inclusions for the coproduct).

Example 8.3.6 (Products and coproducts in $\mathbf{Vect}_k$)

In $\mathbf{Vect}_k$ (and also in $\mathbf{FinVect}_k$), both the product and the coproduct of V and W are realized by the **direct sum** $V \oplus W$.

As a product: Define $\pi_1 : V \oplus W \to V$ by $\pi_1(v, w) = v$ and $\pi_2 : V \oplus W \to W$ by $\pi_2(v, w) = w$. Given linear maps $\alpha : T \to V$ and $\beta : T \to W$, define $\varphi : T \to V \oplus W$ by $\varphi(t) = (\alpha(t), \beta(t))$. It is linear, and it is the unique map with $\pi_1 \circ \varphi = \alpha$ and $\pi_2 \circ \varphi = \beta$, since these two conditions specify both components of $\varphi(t)$.
As a coproduct: Define $\iota_1 : V \to V \oplus W$ by $\iota_1(v) = (v, 0)$ and $\iota_2 : W \to V \oplus W$ by $\iota_2(w) = (0, w)$. Given linear maps $\alpha : V \to T$ and $\beta : W \to T$, define $\varphi : V \oplus W \to T$ by $\varphi(v, w) = \alpha(v) + \beta(w)$. It is linear and satisfies $\varphi \circ \iota_1 = \alpha$ and $\varphi \circ \iota_2 = \beta$. It is the only such map: any linear φ' with these properties satisfies

$$\varphi'(v, w) = \varphi'(\iota_1(v) + \iota_2(w)) = \alpha(v) + \beta(w).$$

NOTE The coincidence of products and coproducts in $\mathbf{Vect}_k$ (and **Ab**) reflects the additive structure of these categories: morphisms can be added, which is what the formula $\alpha(v)+\beta(w)$ uses. In categories without such structure (like **Set** or **Grp**), products and coproducts are genuinely different constructions.

Exercise 8.3.7 (Products and coproducts)

(1) Verify explicitly that the Cartesian product satisfies the universal property of products in **Set**.
(2) Show that in **Ab**, the coproduct of $\mathbb{Z}$ and $\mathbb{Z}$ is $\mathbb{Z}^2$, while in **Grp**, it is the free group F_2 on two generators.
(3) Let V and W be finite-dimensional vector spaces over k. Show that $\dim(V \oplus W) = \dim V + \dim W$ using only the universal property of $V \oplus W$ as a coproduct (not by direct computation).

Solution. **(1)** Let A, B be sets and let T be any set with functions $\alpha : T \to A$ and $\beta : T \to B$. We must show there exists a unique function $\varphi : T \to A \times B$ with $\pi_1 \circ \varphi = \alpha$ and $\pi_2 \circ \varphi = \beta$.
Existence: Define $\varphi(t) = (\alpha(t), \beta(t))$. Then $\pi_1(\varphi(t)) = \alpha(t)$ and $\pi_2(\varphi(t)) = \beta(t)$.
Uniqueness: If $\psi : T \to A \times B$ satisfies $\pi_1 \circ \psi = \alpha$ and $\pi_2 \circ \psi = \beta$, then for any $t \in T$, $\psi(t) = (\pi_1(\psi(t)), \pi_2(\psi(t))) = (\alpha(t), \beta(t)) = \varphi(t)$.
(2) In **Ab**: take $S = \mathbb{Z}^2$ with $\iota_1(n) = (n, 0)$ and $\iota_2(n) = (0, n)$. Let A be an abelian group and $\alpha, \beta : \mathbb{Z} \to A$ homomorphisms. Define $\varphi : \mathbb{Z}^2 \to A$ by $\varphi(m, n) = m\,\alpha(1) + n\,\beta(1)$. This is a homomorphism because A is abelian

(the terms can be rearranged), and $\varphi(\iota_1(m)) = m\,\alpha(1) = \alpha(m)$, $\varphi(\iota_2(n)) = \beta(n)$. It is unique: any homomorphism φ' with $\varphi' \circ \iota_1 = \alpha$ and $\varphi' \circ \iota_2 = \beta$ satisfies $\varphi'(m,n) = \varphi'(m,0) + \varphi'(0,n) = \alpha(m) + \beta(n) = \varphi(m,n)$.

In **Grp**: the coproduct must accommodate non-commuting images. A homomorphism $\alpha : \mathbb{Z} \to G$ is determined by $g = \alpha(1)$, since $\alpha(n) = g^n$, and every $g \in G$ arises this way. Let F_2 be the free group on $\{a, b\}$, with $\iota_1(n) = a^n$ and $\iota_2(n) = b^n$. Given $\alpha, \beta : \mathbb{Z} \to G$, put $g = \alpha(1)$ and $h = \beta(1)$. By the universal property of free groups (Section 3.6), there is a unique homomorphism $\varphi : F_2 \to G$ with $\varphi(a) = g$ and $\varphi(b) = h$. Then $\varphi(\iota_1(n)) = g^n = \alpha(n)$ and $\varphi(\iota_2(n)) = h^n = \beta(n)$. Conversely, any φ' with $\varphi' \circ \iota_1 = \alpha$ and $\varphi' \circ \iota_2 = \beta$ has $\varphi'(a) = g$ and $\varphi'(b) = h$, so $\varphi' = \varphi$. Hence (F_2, ι_1, ι_2) is a coproduct of $\mathbb{Z}$ and $\mathbb{Z}$ in **Grp**. (The group $\mathbb{Z}^2$ is not, since g and h need not commute.)

(3) We use the following characterization of bases, which involves only maps out of U.

Claim. Let U be a vector space and $(u_s)_{s \in S}$ a family of vectors in U. Then (u_s) is a basis of U consisting of distinct vectors iff for every vector space T and every function $f : S \to T$ there is a unique linear map $L : U \to T$ with $L(u_s) = f(s)$ for all $s \in S$.

Proof of the claim. If (u_s) is a basis of distinct vectors, this is the familiar fact that a linear map can be prescribed arbitrarily on a basis, and is determined by these values. Conversely, assume the property. *Spanning:* let $U' = \operatorname{span}(u_s)$, and compare the quotient map $q : U \to U/U'$ (Definition 7.1.1) with the zero map $U \to U/U'$; both send every u_s to 0, so by uniqueness $q = 0$, i.e., $U' = U$. *Independence and distinctness:* let T be a vector space with a basis $(\delta_s)_{s \in S}$ (for instance the space of finitely supported functions $S \to k$), and let L be the linear map with $L(u_s) = \delta_s$. If $\sum_s c_s u_s = 0$ (a finite sum), applying L gives $\sum_s c_s \delta_s = 0$, so all $c_s = 0$; and $u_s = u_{s'}$ forces $\delta_s = \delta_{s'}$, so $s = s'$. This proves the claim.

Now let $(v_i)_{i \in I}$ be a basis of V and $(w_j)_{j \in J}$ a basis of W, with $|I| = \dim V$ and $|J| = \dim W$. Consider the family of $|I| + |J|$ vectors of $V \oplus W$ indexed by the disjoint union $I \sqcup J$: $\iota_1(v_i)$ for $i \in I$ and $\iota_2(w_j)$ for $j \in J$. Let T be any vector space and $f : I \sqcup J \to T$ any function. By the claim for V and W, there are unique linear maps $\alpha : V \to T$ and $\beta : W \to T$ with $\alpha(v_i) = f(i)$ and $\beta(w_j) = f(j)$. By the universal property of the coproduct, there is a unique linear $L : V \oplus W \to T$ with $L \circ \iota_1 = \alpha$ and $L \circ \iota_2 = \beta$. Thus $L(\iota_1(v_i)) = f(i)$ and $L(\iota_2(w_j)) = f(j)$. Any other linear L' with these values satisfies $L' \circ \iota_1 = \alpha$ and $L' \circ \iota_2 = \beta$ (by uniqueness of α and β), hence $L' = L$. So the family has

the property in the claim, and it is a basis of $V \oplus W$ consisting of $|I| + |J|$ distinct vectors. Therefore $\dim(V \oplus W) = \dim V + \dim W$.

NOTE The argument in (3) illustrates a general principle: universal properties translate into bijections of Hom-sets. For instance, the universal property of the coproduct says that

$$\mathrm{Hom}(V \oplus W, T) \to \mathrm{Hom}(V, T) \times \mathrm{Hom}(W, T), \qquad L \mapsto (L \circ \iota_1, L \circ \iota_2),$$

is a bijection (it is even linear): the contravariant functor $\mathrm{Hom}(\cdot, T)$ of Example 8.2.6 turns the coproduct $V \oplus W$ into a product.

CHAPTER 9

Inner Products, Bilinear Forms, and Tensors

So far we have studied vector spaces equipped only with their linear structure. Adding geometric notions—length, angle, orthogonality—requires additional data: an *inner product*. This structure brings abstract linear algebra closer to Euclidean geometry, while revealing that many geometric intuitions extend far beyond $\mathbb{R}^n$.

We begin with the general notion of a bilinear form, then specialize to the symmetric positive-definite case (inner products over $\mathbb{R}$) and its complex analogue (Hermitian inner products over $\mathbb{C}$). We then study the operators adapted to an inner product—adjoints and self-adjoint, normal, orthogonal, and unitary operators—and prove the spectral theorems. Next we return to bilinear forms without any positivity: symmetric forms, which we classify over $\mathbb{C}$ and $\mathbb{R}$, and skew-symmetric (symplectic) forms. The chapter ends with tensor products and multilinear algebra, which turn bilinear and multilinear maps into linear ones.

9.1 Bilinear Forms and Inner Products

9.1.1 Bilinear Forms

Before defining inner products, we study the broader class of bilinear forms. This generality clarifies which properties of inner products come from bilinearity alone and which require symmetry or positivity.

Definition 9.1.1 (Bilinear form)

Let V be a vector space over a field k. A ***bilinear form*** on V is a map $b : V \times V \to k$ that is linear in each variable:

(i) $b(\lambda v, w) = \lambda b(v, w) = b(v, \lambda w)$ for all $\lambda \in k, v, w \in V$.

(ii) $b(u + v, w) = b(u, w) + b(v, w)$ for all $u, v, w \in V$.
(iii) $b(u, v + w) = b(u, v) + b(u, w)$ for all $u, v, w \in V$.

NOTE A bilinear form is *not* a linear map $V \times V \to k$. If we view $V \times V$ as a vector space (the external direct sum $V \oplus V$), a linear map would satisfy $b(\lambda(v, w)) = \lambda b(v, w)$, i.e., $b(\lambda v, \lambda w) = \lambda b(v, w)$. But bilinearity gives $b(\lambda v, \lambda w) = \lambda^2 b(v, w)$. The correct domain for linearizing bilinear maps is the tensor product $V \otimes V$, constructed in Section 9.8.

Definition 9.1.2 (Symmetric, skew-symmetric, and alternating forms)

A bilinear form b on V is:

- ***Symmetric*** if $b(v, w) = b(w, v)$ for all $v, w \in V$.
- ***Skew-symmetric*** if $b(v, w) = -b(w, v)$ for all $v, w \in V$.
- ***Alternating*** if $b(v, v) = 0$ for all $v \in V$.

NOTE Every alternating form is skew-symmetric: expanding $0 = b(v + w, v + w) = b(v, v) + b(v, w) + b(w, v) + b(w, w)$ gives $b(v, w) = -b(w, v)$. Conversely, if $\operatorname{char} k \neq 2$, a skew-symmetric form is alternating, since $b(v, v) = -b(v, v)$ gives $2b(v, v) = 0$. In characteristic 2 the two notions differ: there $-1 = 1$, so skew-symmetric means symmetric, and the dot product on k^2 is skew-symmetric but not alternating ($b(e_1, e_1) = 1$).

Example 9.1.3 (The standard dot product)

On k^n, the map $b(v, w) = \sum_{i=1}^{n} v_i w_i$ is a symmetric bilinear form. This is the "dot product" familiar from calculus.

Example 9.1.4 (A skew-symmetric form)

On k^2, define $b((x_1, x_2), (y_1, y_2)) = x_1 y_2 - x_2 y_1$. This equals $\det \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \end{pmatrix}$ and is alternating, hence skew-symmetric. Over $\mathbb{R}$, it computes the signed area of the parallelogram spanned by the two vectors.

Matrix Representation

Let V be finite-dimensional with a basis $\{e_1, \dots, e_n\}$. A bilinear form b is determined by the values $a_{ij} = b(e_i, e_j)$: by bilinearity,

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

Writing $x = (x_1, \dots, x_n)^t$ and $y = (y_1, \dots, y_n)^t$ as column vectors, this is

$$b(x, y) = x^t A y,$$

where $A = (a_{ij})$ is the ***matrix of*** b with respect to the chosen basis. Conversely, every $n \times n$ matrix A defines a bilinear form by this formula, so the values a_{ij} can be chosen freely.

Proposition 9.1.5 (Symmetry and the matrix)

(i) b is symmetric if and only if $A = A^t$.
(ii) b is skew-symmetric if and only if $A = -A^t$.

Proof. We have $b(v, w) = v^t A w$ and $b(w, v) = w^t A v = (w^t A v)^t = v^t A^t w$, since a 1×1 matrix equals its transpose. So $b(v, w) = b(w, v)$ for all v, w iff $v^t A w = v^t A^t w$ for all v, w, iff $A = A^t$ (take $v = e_i$, $w = e_j$). The skew-symmetric case is the same with a sign. □

The matrix of a bilinear form depends on the basis, and it changes differently from the matrix of an operator.

Proposition 9.1.6 (Change of basis for bilinear forms)

Let $\mathcal{B} = (e_1, \dots, e_n)$ and $\mathcal{B}' = (e'_1, \dots, e'_n)$ be bases of V, and let $P = (p_{ij})$ be the change-of-basis matrix, so that $e'_j = \sum_i p_{ij} e_i$. If A and A' are the matrices of a bilinear form b in the bases $\mathcal{B}$ and $\mathcal{B}'$, then

$$A' = P^t A P.$$

Proof. $a'_{kl} = b(e'_k, e'_l) = b\left(\sum_i p_{ik} e_i, \sum_j p_{jl} e_j\right) = \sum_{i,j} p_{ik} \, a_{ij} \, p_{jl} = (P^t A P)_{kl}$. □

NOTE Matrices related by $A' = P^t A P$ with P invertible are called ***congruent***. Compare with $A' = P^{-1} A P$ for operators (Corollary 5.3.4): the two notions agree when $P^t = P^{-1}$, which will be the case for change of basis between

orthonormal bases (Section 9.5). Since P is invertible, congruent matrices have the same rank, and $\det A' = \det(P)^2 \det A$.

Bilinear Forms and the Dual Space

There is a natural connection between bilinear forms and linear maps to the dual space.

Proposition 9.1.7 (The map $\varphi_b : V \to V^*$)
Given a bilinear form $b : V \times V \to k$, define $\varphi_b : V \to V^*$ by

$$\varphi_b(v) = b(v, \cdot) \in V^*,$$

i.e., $\varphi_b(v)$ is the linear functional $w \mapsto b(v, w)$. Then φ_b is a linear map.

Proof. For $v_1, v_2 \in V$ and $\lambda \in k$, and every $w \in V$:

$$\varphi_b(v_1+\lambda v_2)(w) = b(v_1+\lambda v_2, w) = b(v_1, w)+\lambda b(v_2, w) = (\varphi_b(v_1)+\lambda\varphi_b(v_2))(w).$$

So $\varphi_b(v_1 + \lambda v_2) = \varphi_b(v_1) + \lambda\varphi_b(v_2)$. □

Theorem 9.1.8 (Bilinear forms and $\mathrm{Hom}(V, V^*)$)
Let $B(V)$ denote the vector space of bilinear forms on V. The map $b \mapsto \varphi_b$ is an isomorphism of vector spaces

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

Its inverse sends $\varphi \in \mathrm{Hom}(V, V^*)$ to the bilinear form $b(v, w) = \varphi(v)(w)$.

Proof. The map is linear:

$$\varphi_{b+\lambda b'}(v)(w) = b(v, w) + \lambda b'(v, w) = (\varphi_b + \lambda\varphi_{b'})(v)(w).$$

Given $\varphi \in \mathrm{Hom}(V, V^*)$, the map $b_\varphi(v, w) = \varphi(v)(w)$ is linear in v because φ is linear, and linear in w because $\varphi(v) \in V^*$; so $b_\varphi \in B(V)$. By construction $\varphi_{b_\varphi} = \varphi$ and $b_{\varphi_b} = b$, so the two maps are inverse to each other. □

NOTE In the language of Chapter 8, this isomorphism is natural. A linear map $f : V \to W$ induces maps in the opposite direction,

$$\begin{aligned} B(W) &\to B(V), & b &\mapsto b(f(\cdot), f(\cdot)), \\ \mathrm{Hom}(W, W^*) &\to \mathrm{Hom}(V, V^*), & \varphi &\mapsto f^t \circ \varphi \circ f, \end{aligned}$$

which make $V \mapsto B(V)$ and $V \mapsto \mathrm{Hom}(V, V^*)$ contravariant functors on $\mathbf{Vect}_k$ (Definition 8.2.5). The isomorphisms $b \mapsto \varphi_b$ form a natural transformation between them (Definition 8.2.9): for $b \in B(W)$ and $v, v' \in V$,

$$\varphi_{b(f(\cdot),f(\cdot))}(v)(v') = b(f(v), f(v')) = \varphi_b(f(v))(f(v')) = \big(f^t \circ \varphi_b \circ f\big)(v)(v').$$

Definition 9.1.9 (Rank and nondegeneracy)

Let V be finite-dimensional. The ***rank*** of a bilinear form b is $\mathrm{rank}(\varphi_b) = \dim(\mathrm{Im}(\varphi_b))$. The form b is ***nondegenerate*** if φ_b is an isomorphism.

NOTE Let $\{e_1, \dots, e_n\}$ be a basis of V with dual basis $\{e_1^*, \dots, e_n^*\}$, and let A be the matrix of b. Then $\varphi_b(e_j) = b(e_j, \cdot) = \sum_i b(e_j, e_i)\, e_i^*$, so the matrix of φ_b in these bases is A^t. Hence $\mathrm{rank}\, b = \mathrm{rank}\, A^t = \mathrm{rank}\, A$, and b is nondegenerate iff A is invertible. Since $\dim V = \dim V^*$, φ_b is an isomorphism iff it is injective, so nondegeneracy means: if $b(v, w) = 0$ for all $w \in V$, then $v = 0$. As A is invertible iff A^t is, this is equivalent to the same condition with the roles of the two variables exchanged.

Proposition 9.1.10 (Dimension of the space of bilinear forms)

If $\dim V = n$, then $\dim B(V) = n^2$.

Proof. We have $B(V) \cong \mathrm{Hom}(V, V^*)$ by Theorem 9.1.8, and $\dim \mathrm{Hom}(V, V^*) = n \cdot n = n^2$ since $\dim V^* = n$. Alternatively, $b \mapsto A$ is an isomorphism from $B(V)$ to the space of $n \times n$ matrices, since the entries of A can be chosen freely. □

9.1.2 Inner Products over the Reals

To do geometry, we need more than bilinearity. The key additional properties are symmetry (so that $b(v, w) = b(w, v)$) and positive-definiteness (so that "length" is always nonnegative).

Definition 9.1.11 (Inner product (real case))

Let V be a vector space over $\mathbb{R}$. An ***inner product*** on V is a bilinear form $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ satisfying:

(i) **Symmetry:** $\langle v, w \rangle = \langle w, v \rangle$ for all $v, w \in V$.

(ii) **Positive-definiteness:** $\langle v, v \rangle \geqslant 0$ for all $v \in V$, with equality iff $v = 0$.

A real vector space equipped with an inner product is called a (real) ***inner product space***.

NOTE Positive-definiteness only makes sense over an *ordered* field, where "$\geqslant 0$" is meaningful. In practice, this means $\mathbb{R}$. We cannot directly define inner products over $\mathbb{C}$ or finite fields using this definition; we will see the complex workaround shortly.

Example 9.1.12 (The standard inner product on $\mathbb{R}^n$)

The dot product

$$\langle v, w\rangle = \sum_{i=1}^{n} v_i w_i = v^t w$$

is an inner product. Symmetry is clear, and $\langle v, v\rangle = \sum v_i^2 \geqslant 0$ with equality iff all $v_i = 0$.

Example 9.1.13 (A weighted inner product)

On $\mathbb{R}^n$, for constants $c_1, \dots, c_n > 0$, define

$$\langle v, w\rangle = \sum_{i=1}^{n} c_i v_i w_i.$$

This is an inner product: it is clearly bilinear and symmetric, and $\langle v, v\rangle = \sum c_i v_i^2 \geqslant 0$, with equality iff all $v_i = 0$ because every $c_i > 0$. Different weights emphasize different coordinates.

Proposition 9.1.14 (Inner products are nondegenerate)

If $\langle \cdot, \cdot \rangle$ is an inner product on V, the map $V \to V^*$, $v \mapsto \langle v, \cdot\rangle$, is injective. In particular, if V is finite-dimensional, every inner product on V is nondegenerate.

Proof. Suppose $\langle v, w\rangle = 0$ for all $w \in V$. Taking $w = v$, we get $\langle v, v\rangle = 0$. By positive-definiteness, $v = 0$. □

NOTE The converse is false: a nondegenerate symmetric bilinear form need not be positive-definite. For example, the Minkowski form $b(v, w) =$

$v_1w_1 - v_2w_2 - v_3w_3 - v_4w_4$ on $\mathbb{R}^4$ is nondegenerate and symmetric, but $b(v,v)$ can be negative. Such forms are classified in Section 9.6.

9.1.3 Hermitian Inner Products over the Complex Numbers

Why can't we directly use the same definition over $\mathbb{C}$? If b is a bilinear form on a complex vector space, then $b(iv, iv) = i^2 b(v,v) = -b(v,v)$. So if $b(v,v) > 0$, then $b(iv, iv) < 0$: bilinearity and positive-definiteness are incompatible over $\mathbb{C}$.

The solution is to weaken bilinearity: we require linearity in one variable and only *conjugate*-linearity in the other.

Definition 9.1.15 (Sesquilinear form)

A map $H : V \times V \to \mathbb{C}$ on a complex vector space V is ***sesquilinear*** if, for all $u, v, w \in V$ and $\lambda \in \mathbb{C}$:

(i) $H(u+v, w) = H(u,w) + H(v,w)$ and $H(u, v+w) = H(u,v) + H(u,w)$.

(ii) $H(\lambda u, v) = \overline{\lambda} H(u,v)$.

(iii) $H(u, \lambda v) = \lambda H(u,v)$.

That is, H is conjugate-linear in the first argument and linear in the second.

NOTE This is the convention of Artin (and of most physics texts). Some authors, including Axler, use the opposite convention: linear in the first argument and conjugate-linear in the second. The two conventions are exchanged by swapping the arguments, $H(u,v) \leftrightarrow H(v,u)$, so every formula in one convention has a counterpart in the other; but formulas from the two conventions must not be mixed.

Definition 9.1.16 (Hermitian form)

A sesquilinear form H on a complex vector space V is ***Hermitian*** (or ***conjugate-symmetric***) if

$$H(u,v) = \overline{H(v,u)}$$

for all $u, v \in V$.

NOTE Conjugate-symmetry implies $H(v,v) = \overline{H(v,v)}$, so $H(v,v) \in \mathbb{R}$ for all v. This makes positive-definiteness meaningful.

Definition 9.1.17 (Hermitian inner product)

A ***Hermitian inner product*** on a complex vector space V is a Hermitian form $H : V \times V \to \mathbb{C}$ that is **positive-definite**: $H(v, v) \geqslant 0$ for all v, with equality iff $v = 0$. A complex vector space with a Hermitian inner product is a (complex) ***inner product space***.

Example 9.1.18 (The standard Hermitian inner product on $\mathbb{C}^n$)

Define

$$H(z, w) = \sum_{i=1}^{n} \overline{z_i}\, w_i = z^* w,$$

where $z^* = \overline{z}^{\,t}$ is the conjugate transpose (a row vector). Then H is sesquilinear, and:

- $H(w, z) = \sum \overline{w_i}\, z_i = \overline{\sum \overline{z_i}\, w_i} = \overline{H(z, w)}$.
- $H(z, z) = \sum |z_i|^2 \geqslant 0$, with equality iff $z = 0$.

NOTE If H is a Hermitian form on a finite-dimensional space with basis $\{e_1, \dots, e_n\}$, let $h_{ij} = H(e_i, e_j)$ and $A = (h_{ij})$. Then

$$H\left(\sum_i x_i e_i, \sum_j y_j e_j\right) = \sum_{i,j} \overline{x_i}\, y_j\, h_{ij} = x^* A y,$$

and conjugate-symmetry says $h_{ji} = \overline{h_{ij}}$, i.e., $A^* = A$. A matrix with $A^* = A$ is called ***Hermitian***; conversely, every Hermitian matrix A defines a Hermitian form $x^* A y$ on $\mathbb{C}^n$.

Convention. From now on, an ***inner product space*** is either a real vector space with an inner product or a complex vector space with a Hermitian inner product. We write $\langle \cdot, \cdot \rangle$ in both cases and let $k \in \{\mathbb{R}, \mathbb{C}\}$ be the scalar field. Over $\mathbb{R}$, complex conjugation is the identity, so every statement below specializes to the real case by ignoring the bars. In both cases,

$$\langle \lambda u, v \rangle = \overline{\lambda} \langle u, v \rangle, \qquad \langle u, \lambda v \rangle = \lambda \langle u, v \rangle, \qquad \langle u, v \rangle = \overline{\langle v, u \rangle}.$$

The map $v \mapsto H(v, \cdot)$ is no longer linear, and we need a version of rank–nullity for such maps.

Lemma 9.1.19 (Conjugate-linear maps)

Let V and U be complex vector spaces. A map $f : V \to U$ is ***conjugate-linear*** if $f(u+v) = f(u) + f(v)$ and $f(\lambda v) = \overline{\lambda} f(v)$ for all $u, v \in V$, $\lambda \in \mathbb{C}$. If f is conjugate-linear, then $\ker f$ and $\operatorname{Im} f$ are subspaces, and if V is finite-dimensional,

$$\dim V = \dim \ker f + \dim \operatorname{Im} f.$$

A composite of two conjugate-linear maps is linear, and a composite of a linear map and a conjugate-linear map (in either order) is conjugate-linear.

Proof. Let $\overline{U}$ be the set U with the same addition and with the new scalar multiplication $\lambda \star u = \overline{\lambda} u$. Since complex conjugation is a field automorphism of $\mathbb{C}$ ($\overline{\lambda+\mu} = \overline{\lambda} + \overline{\mu}$, $\overline{\lambda\mu} = \overline{\lambda}\,\overline{\mu}$, $\overline{1} = 1$), the vector space axioms for $\overline{U}$ follow from those of U. A subset of U is closed under all operations $u \mapsto \lambda \star u$ iff it is closed under all $u \mapsto \overline{\lambda} u$, i.e., under all scalar multiplications; so U and $\overline{U}$ have the same subspaces. A list (u_i) is a basis of a subspace in U iff it is one in $\overline{U}$, because $\sum \lambda_i \star u_i = \sum \overline{\lambda_i} u_i$ and $\lambda \mapsto \overline{\lambda}$ is a bijection; so dimensions agree. Finally, $f(\lambda v) = \overline{\lambda} f(v) = \lambda \star f(v)$, so $f : V \to \overline{U}$ is linear, and the claims about $\ker f$ and $\operatorname{Im} f$ follow from rank–nullity (Theorem 6.1.12). The statements about composites are immediate: two conjugations cancel, $\overline{\overline{\lambda}} = \lambda$. □

Proposition 9.1.20 (The map $\varphi_H : V \to V^*$)

Let H be a Hermitian form on a complex vector space V. For $v \in V$, the map $H(v, \cdot) : w \mapsto H(v, w)$ is linear, so $\varphi_H(v) = H(v, \cdot)$ lies in V^*. The map $\varphi_H : V \to V^*$ is **conjugate-linear**:

$$\varphi_H(\lambda v) = \overline{\lambda}\, \varphi_H(v).$$

We call H ***nondegenerate*** if φ_H is injective, i.e., if $H(v, w) = 0$ for all w implies $v = 0$. A Hermitian inner product is nondegenerate, and if V is finite-dimensional and H is nondegenerate, then φ_H is bijective.

Proof. The functional $H(v, \cdot)$ is linear because H is linear in its second argument, and for all w, $\varphi_H(\lambda v)(w) = H(\lambda v, w) = \overline{\lambda} H(v, w)$. If H is positive-definite and $H(v, w) = 0$ for all w, then $H(v, v) = 0$, so $v = 0$. If V is finite-dimensional and φ_H is injective, Lemma 9.1.19 gives $\dim \operatorname{Im} \varphi_H = \dim V = \dim V^*$, so φ_H is also surjective. □

Combining the real and complex cases, we obtain the fundamental fact that every linear functional on a finite-dimensional inner product space is "inner product with a vector."

Theorem 9.1.21 (Riesz representation)

Let V be a finite-dimensional inner product space. For every $\ell \in V^*$ there is a unique $u \in V$ such that

$$\ell(w) = \langle u, w \rangle \quad \text{for all } w \in V.$$

Equivalently, the map $\Phi : V \to V^*$, $\Phi(u) = \langle u, \cdot \rangle$, is a bijection. It is linear if $k = \mathbb{R}$ and conjugate-linear if $k = \mathbb{C}$.

Proof. Over $\mathbb{R}$, $\Phi = \varphi_b$ for the bilinear form $b = \langle \cdot, \cdot \rangle$, which is nondegenerate by Proposition 9.1.14, so Φ is an isomorphism. Over $\mathbb{C}$, $\Phi = \varphi_H$ is a conjugate-linear bijection by Proposition 9.1.20. □

9.1.4 Norms and the Cauchy–Schwarz Inequality

An inner product induces a notion of length via the norm.

Definition 9.1.22 (Norm)

Let V be an inner product space. The ***norm*** of a vector $v \in V$ is

$$\|v\| = \sqrt{\langle v, v \rangle}.$$

This is well-defined since $\langle v, v \rangle$ is a nonnegative real number.

Definition 9.1.23 (Orthogonality)

Vectors $v, w \in V$ are ***orthogonal***, written $v \perp w$, if $\langle v, w \rangle = 0$ (equivalently, $\langle w, v \rangle = \overline{\langle v, w \rangle} = 0$).

Proposition 9.1.24 (Basic properties of the norm)

(i) $\|v\| \geqslant 0$, with equality iff $v = 0$.
(ii) $\|\lambda v\| = |\lambda| \|v\|$ for $\lambda \in k$.
(iii) (Pythagorean theorem) If $v \perp w$, then $\|v + w\|^2 = \|v\|^2 + \|w\|^2$.

Proof. (i) is positive-definiteness. (ii) $\langle \lambda v, \lambda v \rangle = \overline{\lambda}\lambda \langle v, v \rangle = |\lambda|^2 \|v\|^2$. For (iii):

$$\|v + w\|^2 = \langle v + w, v + w \rangle = \|v\|^2 + \langle v, w \rangle + \langle w, v \rangle + \|w\|^2 = \|v\|^2 + \|w\|^2,$$

since $\langle v, w \rangle = 0 = \langle w, v \rangle$. □

The Pythagorean theorem suggests that orthogonal vectors behave like the legs of a right triangle. More generally, since $\langle v, w\rangle + \langle w, v\rangle = 2\,\mathrm{Re}\langle v, w\rangle$,

$$\|v - w\|^2 = \|v\|^2 + \|w\|^2 - 2\,\mathrm{Re}\langle v, w\rangle.$$

In a real inner product space, comparing with the law of cosines suggests defining the angle θ between nonzero vectors by $\cos\theta = \langle v, w\rangle/(\|v\|\|w\|)$. For this we first need to know that $|\langle v, w\rangle| \leqslant \|v\|\|w\|$, so that the ratio lies in $[-1, 1]$.

Theorem 9.1.25 (Cauchy–Schwarz inequality)

For all u, v in an inner product space,

$$|\langle u, v\rangle| \leqslant \|u\|\|v\|,$$

with equality iff u and v are linearly dependent.

Proof. If $u = 0$, both sides are 0 and u, v are dependent. Assume $u \neq 0$. Replacing u by $u/\|u\|$ multiplies both sides by $1/\|u\|$ and does not change linear dependence, so we may assume $\|u\| = 1$.

Decompose v along u:

$$v = v_1 + v_2, \qquad v_1 = \langle u, v\rangle u, \qquad v_2 = v - \langle u, v\rangle u.$$

Then $\langle u, v_2\rangle = \langle u, v\rangle - \langle u, v\rangle\langle u, u\rangle = 0$, so $v_2 \perp u$ and hence $v_2 \perp v_1$. By the Pythagorean theorem,

$$\|v\|^2 = \|v_1\|^2 + \|v_2\|^2 \geqslant \|v_1\|^2 = |\langle u, v\rangle|^2\|u\|^2 = |\langle u, v\rangle|^2.$$

Taking square roots gives $|\langle u, v\rangle| \leqslant \|v\| = \|u\|\|v\|$.

Equality holds iff $v_2 = 0$, i.e., iff $v \in \mathrm{span}(u)$, i.e. (as $u \neq 0$) iff u and v are linearly dependent. □

Definition 9.1.26 (Angle between vectors)

For nonzero vectors v, w in a real inner product space, the ***angle*** $\theta \in [0, \pi]$ between them is defined by

$$\cos\theta = \frac{\langle v, w\rangle}{\|v\|\|w\|}.$$

The Cauchy–Schwarz inequality ensures this ratio lies in $[-1, 1]$.

Corollary 9.1.27 (Triangle inequality)
For all u, v in an inner product space,

$$\|u + v\| \leqslant \|u\| + \|v\|.$$

Proof.

$$\begin{aligned}\|u+v\|^2 &= \|u\|^2 + 2\,\mathrm{Re}\langle u, v\rangle + \|v\|^2 \\ &\leqslant \|u\|^2 + 2|\langle u, v\rangle| + \|v\|^2 \\ &\leqslant \|u\|^2 + 2\|u\|\|v\| + \|v\|^2 = (\|u\| + \|v\|)^2.\end{aligned}$$

□

9.1.5 Orthogonal Complements

Orthogonality makes sense for any bilinear or Hermitian form, not only for inner products.

Definition 9.1.28 (Orthogonal complement)

Let b be a bilinear form on a vector space V, or a Hermitian form on a complex vector space V, and let $S \subseteq V$ be a subset. The ***orthogonal complement*** of S is

$$S^\perp = \{v \in V : b(v, w) = 0 \text{ for all } w \in S\}.$$

For an inner product, $S^\perp = \{v \in V : \langle v, w\rangle = 0 \text{ for all } w \in S\}$.

Proposition 9.1.29 (Basic properties of orthogonal complements)

(i) $S^\perp$ is a subspace of V (even if S is not).
(ii) $S \subseteq T \implies T^\perp \subseteq S^\perp$.
(iii) If b is symmetric, skew-symmetric, or Hermitian, then $S \subseteq (S^\perp)^\perp$.
(iv) $\{0\}^\perp = V$, and $V^\perp = \ker \varphi_b$. For an inner product, $V^\perp = \{0\}$.

Proof. (i) If $v_1, v_2 \in S^\perp$ and λ is a scalar, then for any $w \in S$,

$$b(v_1 + \lambda v_2, w) = b(v_1, w) + \lambda' b(v_2, w) = 0,$$

where $\lambda' = \lambda$ for a bilinear form and $\lambda' = \overline{\lambda}$ for a Hermitian form.

(ii) is immediate. (iii) If $w \in S$ and $v \in S^\perp$, then $b(w, v)$ equals $b(v, w)$, $-b(v, w)$, or $\overline{b(v, w)}$, which is 0 in each case; so $w \in (S^\perp)^\perp$. (iv) The first two statements are the definitions, and the last is Proposition 9.1.14 (or Proposition 9.1.20). □

NOTE For a general bilinear form (neither symmetric nor skew-symmetric), we must distinguish "left-orthogonal" ($b(v,w)=0$ for all $w \in S$) from "right-orthogonal" ($b(w,v)=0$ for all $w \in S$). For symmetric, skew-symmetric, and Hermitian forms these coincide.

The key result is that an inner product space decomposes as $V = S \oplus S^{\perp}$ for any subspace S. We first count dimensions.

Lemma 9.1.30 (Dimension of the orthogonal complement)

Let b be a nondegenerate bilinear form, or a nondegenerate Hermitian form, on a finite-dimensional space V, and let $S \subseteq V$ be a subspace. Then

$$\dim S^{\perp} = \dim V - \dim S.$$

If b is an inner product, then moreover $S \cap S^{\perp} = \{0\}$.

Proof. Consider the composition

$$V \xrightarrow{\varphi_b} V^* \xrightarrow{r} S^*,$$

where $\varphi_b(v) = b(v,\cdot)$ and $r : V^* \to S^*$ is restriction, $r(\ell) = \ell|_S$. The restriction r is the dual map of the inclusion $S \hookrightarrow V$, which is injective, so r is surjective (Corollary 6.2.16). The kernel of $r \circ \varphi_b$ is exactly $S^{\perp}$.

Since φ_b is bijective (by nondegeneracy) and r is surjective, $r \circ \varphi_b$ is surjective. It is linear for a bilinear form and conjugate-linear for a Hermitian form, so in either case rank–nullity (Theorem 6.1.12, or Lemma 9.1.19) gives

$$\dim S^{\perp} = \dim V - \dim S^* = \dim V - \dim S.$$

For the second claim: if $v \in S \cap S^{\perp}$, then $\langle v, v\rangle = 0$, so $v = 0$ by positive-definiteness. □

Theorem 9.1.31 (Orthogonal decomposition)

Let V be a finite-dimensional inner product space and $S \subseteq V$ a subspace. Then

$$V = S \oplus S^{\perp}.$$

Proof. An inner product is nondegenerate (Proposition 9.1.14 or Proposition 9.1.20), so by the lemma, $\dim S + \dim S^{\perp} = \dim V$ and $S \cap S^{\perp} = \{0\}$. Hence

$$\dim(S + S^{\perp}) = \dim S + \dim S^{\perp} - \dim(S \cap S^{\perp}) = \dim V.$$

So $S + S^{\perp} = V$, and the sum is direct since $S \cap S^{\perp} = \{0\}$. □

Corollary 9.1.32 (Double orthogonal complement)

Let b be a nondegenerate symmetric, skew-symmetric, or Hermitian form on a finite-dimensional space V (for instance an inner product). Then $(S^\perp)^\perp = S$ for every subspace $S \subseteq V$.

Proof. By Proposition 9.1.29(iii), $S \subseteq (S^\perp)^\perp$, and by Lemma 9.1.30 applied twice, $\dim(S^\perp)^\perp = \dim V - (\dim V - \dim S) = \dim S$. □

Example 9.1.33 (Standard orthogonal complement in $\mathbb{R}^n$)

Let $V = \mathbb{R}^n$ with the standard inner product. If S is a subspace, then $S^\perp$ is the "usual" orthogonal complement: the vectors perpendicular to every vector in S.
For instance, if $S = \operatorname{span}((1,0,0))$ in $\mathbb{R}^3$, then $S^\perp = \{(0,y,z) : y,z \in \mathbb{R}\}$.

Example 9.1.34 (Skew-symmetric forms can have $S^\perp = S$)

Let $V = k^2$ with the skew-symmetric form $b((x_1,x_2),(y_1,y_2)) = x_1y_2 - x_2y_1$ of Example 9.1.4, which is nondegenerate (its matrix $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ is invertible).
Let S be the line spanned by a nonzero vector v.
For any $w \in k^2$, we have $b(v,w) = 0 \iff \det(v \mid w) = 0 \iff w \in \operatorname{span}(v) = S$.
So $S^\perp = S$. In particular, $S \cap S^\perp = S \neq \{0\}$, and $V \neq S \oplus S^\perp$. This cannot happen for inner products.

9.2 Orthonormal Bases

9.2.1 The Gram–Schmidt Process

In an inner product space, we can ask for bases that respect the geometry: bases whose vectors are mutually orthogonal and have unit length.

Definition 9.2.1 (Orthonormal basis)

Let V be an inner product space. A list $(v_1,\dots,v_m)$ of vectors in V is ***orthonormal*** if

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

Equivalently, $\|v_i\| = 1$ for all i, and $v_i \perp v_j$ for $i \neq j$. An ***orthonormal basis*** is an orthonormal list that is a basis.

NOTE An orthonormal list is linearly independent: if $\sum_i c_i v_i = 0$, then $0 = \langle v_j, \sum_i c_i v_i \rangle = \sum_i c_i \langle v_j, v_i \rangle = c_j$ for each j. So in an n-dimensional space, every orthonormal list of length n is an orthonormal basis.
In an orthonormal basis $\{e_1, \dots, e_n\}$, coordinates are computed by inner products: if $v = \sum x_i e_i$, then $\langle e_j, v \rangle = x_j$, so

$$v = \sum_{i=1}^{n} \langle e_i, v \rangle \, e_i.$$

The inner product also takes a simple form: if $w = \sum y_j e_j$, then

$$\langle v, w \rangle = \sum_{i,j} \overline{x_i} \, y_j \langle e_i, e_j \rangle = \sum_i \overline{x_i} \, y_i.$$

For real spaces this is $x^{t} y$; for complex spaces it is $x^{*} y$.

Theorem 9.2.2 (Existence of orthonormal bases)
Every finite-dimensional inner product space has an orthonormal basis.

We give two proofs: one by induction, one constructive (Gram–Schmidt).

Proof 1 (Induction on dimension). Let $\dim V = n$. If $n = 0$, the empty list is an orthonormal basis. For $n \geqslant 1$:

Pick any nonzero $v \in V$ and set $v_1 = v/\|v\|$, so $\|v_1\| = 1$. Let $S = \operatorname{span}(v_1)$. By the orthogonal decomposition theorem, $V = S \oplus S^{\perp}$, and $\dim S^{\perp} = n - 1$.

The restriction of $\langle \cdot, \cdot \rangle$ to $S^{\perp}$ is still an inner product (the defining properties are inherited by a subspace). By induction, $S^{\perp}$ has an orthonormal basis $\{v_2, \dots, v_n\}$.

Then $(v_1, v_2, \dots, v_n)$ is orthonormal, because $v_1 \perp v_i$ for $i \geqslant 2$ (as $v_i \in S^{\perp}$); being of length n, it is an orthonormal basis of V. $\square$

Proof 2 (Gram–Schmidt). Start with any basis $\{w_1, \dots, w_n\}$ of V. We inductively construct orthonormal vectors $v_1, \dots, v_n$ such that for each j,

$$\operatorname{span}(v_1, \dots, v_j) = \operatorname{span}(w_1, \dots, w_j).$$

Step 1: Set $v_1 = w_1/\|w_1\|$.

Step j**:** Suppose $v_1, \dots, v_{j-1}$ are orthonormal and span the same subspace as $w_1, \dots, w_{j-1}$. Define

$$u_j = w_j - \sum_{i=1}^{j-1} \langle v_i, w_j \rangle \, v_i.$$

This subtracts from w_j its components along each v_i: for $l < j$,

$$\langle v_l, u_j \rangle = \langle v_l, w_j \rangle - \sum_{i=1}^{j-1} \langle v_i, w_j \rangle \langle v_l, v_i \rangle = \langle v_l, w_j \rangle - \langle v_l, w_j \rangle = 0.$$

Since $w_j \notin \operatorname{span}(w_1, \dots, w_{j-1}) = \operatorname{span}(v_1, \dots, v_{j-1})$ (linear independence), we have $u_j \neq 0$. Set

$$v_j = \frac{u_j}{\|u_j\|}.$$

By construction, $v_j \perp v_i$ for $i < j$, and $\|v_j\| = 1$. The span condition holds since $v_j \in \operatorname{span}(v_1, \dots, v_{j-1}, w_j) = \operatorname{span}(w_1, \dots, w_j)$ and $w_j \in \operatorname{span}(v_1, \dots, v_j)$. □

Corollary 9.2.3 (Isomorphism with standard space)

Every n-dimensional real inner product space is isomorphic to $\mathbb{R}^n$ with the standard inner product, as an inner product space (that is, by a linear isomorphism that preserves inner products). Similarly, every n-dimensional complex inner product space is isomorphic to $\mathbb{C}^n$ with the standard Hermitian inner product.

Proof. Let $\{v_1, \dots, v_n\}$ be an orthonormal basis. The coordinate map $\varphi : V \to k^n$ sending $v = \sum a_i v_i$ to $(a_1, \dots, a_n)$ is a linear isomorphism. Moreover, for $w = \sum b_j v_j$,

$$\langle v, w \rangle = \Big\langle \sum_i a_i v_i, \sum_j b_j v_j \Big\rangle = \sum_{i,j} \overline{a_i} \, b_j \, \delta_{ij} = \sum_i \overline{a_i} \, b_i,$$

which is the standard inner product of the coordinate vectors. □

NOTE This says that all n-dimensional inner product spaces over the same field are "the same." The choice of orthonormal basis corresponds to the choice of isomorphism with the standard space. Different orthonormal bases give different coordinate systems, all equally valid.

Example 9.2.4 (Gram–Schmidt in $\mathbb{R}^3$)

Apply Gram–Schmidt to the basis $w_1 = (1,1,0)$, $w_2 = (1,0,1)$, $w_3 = (0,1,1)$ of $\mathbb{R}^3$ with the standard inner product.
Step 1: $\|w_1\| = \sqrt{2}$, so $v_1 = \frac{1}{\sqrt{2}}(1,1,0)$.
Step 2: $\langle v_1, w_2\rangle = \frac{1}{\sqrt{2}}(1+0+0) = \frac{1}{\sqrt{2}}$, so

$$u_2 = w_2 - \frac{1}{\sqrt{2}} v_1 = (1,0,1) - \frac{1}{2}(1,1,0) = \left(\frac{1}{2}, -\frac{1}{2}, 1\right).$$

$\|u_2\|^2 = \frac{1}{4} + \frac{1}{4} + 1 = \frac{3}{2}$, so $v_2 = \sqrt{\frac{2}{3}}\left(\frac{1}{2}, -\frac{1}{2}, 1\right) = \frac{1}{\sqrt{6}}(1,-1,2)$.
Step 3: $\langle v_1, w_3\rangle = \frac{1}{\sqrt{2}}$ and $\langle v_2, w_3\rangle = \frac{1}{\sqrt{6}}(0-1+2) = \frac{1}{\sqrt{6}}$, so

$$u_3 = w_3 - \frac{1}{\sqrt{2}} v_1 - \frac{1}{\sqrt{6}} v_2 = (0,1,1) - \frac{1}{2}(1,1,0) - \frac{1}{6}(1,-1,2) = \left(-\frac{2}{3}, \frac{2}{3}, \frac{2}{3}\right).$$

$\|u_3\|^2 = \frac{4}{3}$, so $v_3 = \frac{\sqrt{3}}{2}\left(-\frac{2}{3}, \frac{2}{3}, \frac{2}{3}\right) = \frac{1}{\sqrt{3}}(-1,1,1)$.
One checks directly that $v_3 \perp v_1$ and $v_3 \perp v_2$.

9.2.2 Orthogonal Projections

The orthogonal decomposition $V = S \oplus S^{\perp}$ gives a canonical way to project onto S. It holds whenever S is finite-dimensional, even if V is not, and an orthonormal basis of S gives an explicit formula.

Proposition 9.2.5 (Projection onto a finite-dimensional subspace)
Let V be an inner product space (not necessarily finite-dimensional) and $S \subseteq V$ a finite-dimensional subspace with orthonormal basis $\{e_1, \dots, e_k\}$. Then $V = S \oplus S^{\perp}$, and the decomposition of $v \in V$ is

$$v = \sum_{i=1}^{k} \langle e_i, v\rangle\, e_i + \Big(v - \sum_{i=1}^{k} \langle e_i, v\rangle\, e_i\Big) \in S \oplus S^{\perp}.$$

Proof. Let $w = \sum_{i=1}^{k} \langle e_i, v\rangle e_i \in S$. For each $j \leqslant k$,

$$\langle e_j, v - w\rangle = \langle e_j, v\rangle - \sum_{i} \langle e_i, v\rangle \langle e_j, e_i\rangle = \langle e_j, v\rangle - \langle e_j, v\rangle = 0.$$

Hence $\langle v - w, e_j\rangle = 0$ for all j, and since the e_j span S, $v - w \in S^{\perp}$. So $V = S + S^{\perp}$, and $S \cap S^{\perp} = \{0\}$ by positive-definiteness. $\square$

Definition 9.2.6 (Orthogonal projection)

Let V be an inner product space and $S \subseteq V$ a subspace with $V = S \oplus S^\perp$ (for instance, S finite-dimensional). The ***orthogonal projection*** onto S is the map $\pi_S : V \to V$ defined by

$$\pi_S(v) = v_S, \quad \text{where } v = v_S + v_{S^\perp} \text{ with } v_S \in S,\ v_{S^\perp} \in S^\perp.$$

It is linear, by uniqueness of the decomposition.

Proposition 9.2.7 (Properties of orthogonal projections)

Let $\pi_S : V \to V$ be the orthogonal projection onto S.

(i) $\mathrm{Im}(\pi_S) = S$ and $\ker(\pi_S) = S^\perp$.
(ii) $\pi_S^2 = \pi_S$ (i.e., π_S is idempotent).
(iii) $\langle \pi_S(u), v\rangle = \langle u, \pi_S(v)\rangle$ for all $u, v \in V$.
(iv) If V is finite-dimensional, then $\pi_{S^\perp} = \mathrm{id} - \pi_S$.

Proof. (i) follows from the definition. (ii) If $v = v_S + v_{S^\perp}$, then $\pi_S(v) = v_S \in S$ has decomposition $v_S + 0$, so $\pi_S(\pi_S(v)) = v_S$.

(iii) Write $u = u_S + u'$ and $v = v_S + v'$ with $u', v' \in S^\perp$. Then

$$\langle \pi_S(u), v\rangle = \langle u_S, v_S + v'\rangle = \langle u_S, v_S\rangle = \langle u_S + u', v_S\rangle = \langle u, \pi_S(v)\rangle.$$

(iv) We have $v = (v - \pi_S(v)) + \pi_S(v)$ with $v - \pi_S(v) \in S^\perp$ and $\pi_S(v) \in S = (S^\perp)^\perp$ (Corollary 9.1.32). This is the decomposition of v in $V = S^\perp \oplus (S^\perp)^\perp$, so $\pi_{S^\perp}(v) = v - \pi_S(v)$. □

NOTE In the language of the next section, (iii) says that π_S is self-adjoint.

Proposition 9.2.8 (Best approximation)

For $v \in V$ and $S \subseteq V$ a finite-dimensional subspace, the projection $\pi_S(v)$ is the unique vector in S minimizing $\|v - s\|$ over $s \in S$.

Proof. For any $s \in S$, write

$$v - s = (v - \pi_S(v)) - (s - \pi_S(v)).$$

Since $v - \pi_S(v) \in S^\perp$ and $s - \pi_S(v) \in S$, these two vectors are orthogonal, and by the Pythagorean theorem

$$\|v - s\|^2 = \|v - \pi_S(v)\|^2 + \|s - \pi_S(v)\|^2 \geqslant \|v - \pi_S(v)\|^2,$$

with equality iff $s = \pi_S(v)$. □

Example 9.2.9 (Fourier series)
Let $V = C^\infty(S^1, \mathbb{C})$ be the space of infinitely differentiable functions $f : \mathbb{R} \to \mathbb{C}$ with period 1 (functions on the circle $S^1 = \mathbb{R}/\mathbb{Z}$), with

$$\langle f, g \rangle = \int_0^1 \overline{f(t)}\, g(t)\, dt.$$

This is a Hermitian inner product ($\langle f, f \rangle = \int_0^1 |f(t)|^2 dt > 0$ for a continuous $f \neq 0$). The functions $f_n(t) = e^{2\pi i n t}$, $n \in \mathbb{Z}$, are orthonormal:

$$\langle f_n, f_m \rangle = \int_0^1 e^{2\pi i (m-n) t}\, dt = \delta_{mn}.$$

They are not a basis of V: their span is the subspace of trigonometric polynomials. But for each N, Proposition 9.2.5 applies to $S_N = \operatorname{span}(f_{-N}, \dots, f_N)$:

$$\pi_{S_N}(f) = \sum_{n=-N}^{N} \langle f_n, f \rangle f_n, \qquad \langle f_n, f \rangle = \int_0^1 e^{-2\pi i n t} f(t)\, dt.$$

The numbers $\langle f_n, f \rangle$ are the Fourier coefficients of f, so the partial sums of the Fourier series of f are orthogonal projections, and by best approximation $\pi_{S_N}(f)$ is the trigonometric polynomial of degree at most N closest to f. Making sense of the full infinite series requires analysis (Hilbert spaces), which we do not pursue.

9.3 The Adjoint

9.3.1 Definition and Existence

Given a linear operator on an inner product space, its *adjoint* is the operator obtained by moving it from one argument of the inner product to the other. Throughout this section V is a finite-dimensional inner product space.

Theorem 9.3.1 (Existence and uniqueness of the adjoint)
Let $T : V \to V$ be a linear operator on a finite-dimensional inner product space. There is a unique map $T^* : V \to V$ such that

$$\langle v, T(w) \rangle = \langle T^*(v), w \rangle \quad \text{for all } v, w \in V,$$

and T^* is linear.

Proof. **Existence.** Fix $v \in V$. The map $w \mapsto \langle v, T(w)\rangle$ is a linear functional on V, being the composite of T with the linear functional $\langle v, \cdot\rangle$. By the Riesz representation theorem, there exists a unique $u \in V$ with $\langle u, w\rangle = \langle v, T(w)\rangle$ for all w. Define $T^*(v) = u$.

Linearity. For $v_1, v_2 \in V$, $\lambda \in k$, and all $w \in V$:

$$\begin{aligned}\langle T^*(v_1 + \lambda v_2), w\rangle &= \langle v_1 + \lambda v_2, T(w)\rangle = \langle v_1, T(w)\rangle + \overline{\lambda}\langle v_2, T(w)\rangle \\ &= \langle T^*(v_1), w\rangle + \overline{\lambda}\langle T^*(v_2), w\rangle = \langle T^*(v_1) + \lambda T^*(v_2), w\rangle.\end{aligned}$$

By the uniqueness in the Riesz representation theorem, $T^*(v_1 + \lambda v_2) = T^*(v_1) + \lambda T^*(v_2)$.

Uniqueness. If S also satisfies $\langle v, T(w)\rangle = \langle S(v), w\rangle$ for all v, w, then $\langle T^*(v) - S(v), w\rangle = 0$ for all w, so $T^*(v) - S(v) \in V^{\perp} = \{0\}$. □

Definition 9.3.2 (Adjoint operator)

Let $T : V \to V$ be a linear operator on a finite-dimensional inner product space. The ***adjoint*** of T is the operator $T^* : V \to V$ of the theorem, characterized by

$$\langle v, T(w)\rangle = \langle T^*(v), w\rangle \quad \text{for all } v, w \in V.$$

Taking complex conjugates of both sides (and renaming the variables), this is equivalent to $\langle T(v), w\rangle = \langle v, T^*(w)\rangle$ for all v, w.

NOTE The adjoint is the dual map transported by the inner product. Let $\Phi : V \to V^*$, $\Phi(v) = \langle v, \cdot\rangle$, be the bijection of the Riesz representation theorem, and let $T^t : V^* \to V^*$, $T^t(\ell) = \ell \circ T$, be the dual map (Definition 6.2.9). The defining property says

$$\Phi(T^*(v)) = \langle T^*(v), \cdot\rangle = \langle v, T(\cdot)\rangle = \Phi(v) \circ T = T^t(\Phi(v)),$$

so

$$T^* = \Phi^{-1} \circ T^t \circ \Phi.$$

Over $\mathbb{R}$, Φ is linear and T^* is simply the transpose T^t read through the isomorphism $V \cong V^*$. Over $\mathbb{C}$, Φ is conjugate-linear, and T^* is the *complex conjugate* of the transpose: in an orthonormal basis $\{e_i\}$ with dual basis $\{e_i^*\}$ we have $\Phi(\sum x_i e_i) = \sum \overline{x_i}\, e_i^*$, so if T has matrix M, then T^t has matrix M^t and T^* has matrix $\overline{M^t}$ (see Theorem 9.3.4 below).

9.3.2 Properties of the Adjoint

Proposition 9.3.3 (Algebraic properties of the adjoint)

Let $S, T : V \to V$ be linear operators on a finite-dimensional inner product space, and $\lambda \in k$. Then:

(i) $(T^*)^* = T$.
(ii) $(S + T)^* = S^* + T^*$.
(iii) $(\lambda T)^* = \overline{\lambda} T^*$.
(iv) $(ST)^* = T^* S^*$.
(v) $\mathrm{id}^* = \mathrm{id}$.
(vi) If T is invertible, then T^* is invertible and $(T^{-1})^* = (T^*)^{-1}$.

Proof. Each follows from the defining equation $\langle v, T(w)\rangle = \langle T^*(v), w\rangle$ and the uniqueness of the adjoint.

(i) $\langle v, T^*(w)\rangle = \overline{\langle T^*(w), v\rangle} = \overline{\langle w, T(v)\rangle} = \langle T(v), w\rangle$. So T satisfies the defining property of $(T^*)^*$.

(ii) $\langle v, (S+T)(w)\rangle = \langle v, S(w)\rangle + \langle v, T(w)\rangle = \langle S^*(v) + T^*(v), w\rangle$.

(iii) $\langle v, \lambda T(w)\rangle = \lambda \langle v, T(w)\rangle = \lambda \langle T^*(v), w\rangle = \langle \overline{\lambda} T^*(v), w\rangle$.

(iv) $\langle v, S(T(w))\rangle = \langle S^*(v), T(w)\rangle = \langle T^*(S^*(v)), w\rangle$.

(v) $\langle v, \mathrm{id}(w)\rangle = \langle \mathrm{id}(v), w\rangle$.

(vi) Applying (iv) and (v) to $TT^{-1} = \mathrm{id} = T^{-1}T$ gives $(T^{-1})^* T^* = \mathrm{id} = T^*(T^{-1})^*$. □

NOTE Property (iv) shows that taking adjoints reverses the order of products, just like taking transposes or inverses. So $T \mapsto T^*$ is a conjugate-linear map $\mathrm{End}(V) \to \mathrm{End}(V)$ that reverses products and is its own inverse.

Matrix of the Adjoint

Theorem 9.3.4 (Matrix of the adjoint)

Let $\{e_1, \dots, e_n\}$ be an orthonormal basis of V, and let $M = \mathcal{M}(T)$ be the matrix of T in this basis. Then the matrix of T^* is

$$\mathcal{M}(T^*) = \begin{cases} M^{\mathrm{t}} & \text{(real inner product)}, \\ M^* = \overline{M}^{\mathrm{t}} & \text{(Hermitian inner product)}, \end{cases}$$

where M^* denotes the conjugate transpose.

Proof. The (i,j)-entry of $\mathcal{M}(T)$ is the coefficient of e_i in $T(e_j)$, which is $\langle e_i, T(e_j)\rangle$ (coordinates in an orthonormal basis). Likewise the (i,j)-entry of $\mathcal{M}(T^*)$ is

$$\langle e_i, T^*(e_j)\rangle = \overline{\langle T^*(e_j), e_i\rangle} = \overline{\langle e_j, T(e_i)\rangle} = \overline{\mathcal{M}(T)_{ji}}.$$

So $\mathcal{M}(T^*)_{ij} = \overline{M_{ji}} = (M^*)_{ij}$; over $\mathbb{R}$ the bar has no effect. □

NOTE The orthonormality of the basis matters: in a general basis, the matrix of T^* is not M^*.

Proposition 9.3.5 (Kernel and image of the adjoint)

For $T : V \to V$:

(i) $\ker(T^*) = \operatorname{Im}(T)^\perp$.

(ii) $\operatorname{Im}(T^*) = \ker(T)^\perp$.

Proof. (i) Since $V^\perp = \{0\}$,

$$v \in \ker(T^*) \iff \langle T^*(v), w\rangle = 0 \ \forall w \iff \langle v, T(w)\rangle = 0 \ \forall w \iff v \in \operatorname{Im}(T)^\perp.$$

(ii) Apply (i) to T^*: $\ker((T^*)^*) = \operatorname{Im}(T^*)^\perp$, i.e., $\ker(T) = \operatorname{Im}(T^*)^\perp$. Taking orthogonal complements and using Corollary 9.1.32: $\ker(T)^\perp = (\operatorname{Im}(T^*)^\perp)^\perp = \operatorname{Im}(T^*)$. □

Corollary 9.3.6 (Rank equality)

$\operatorname{rank}(T) = \operatorname{rank}(T^*)$.

Proof. $\operatorname{rank}(T^*) = \dim \operatorname{Im}(T^*) = \dim \ker(T)^\perp = \dim V - \dim \ker(T) = \operatorname{rank}(T)$, by Lemma 9.1.30 and rank–nullity. □

The following lemma is the key to all the structure theorems below: it lets us split off an invariant subspace and continue on its orthogonal complement.

Lemma 9.3.7 (Invariant subspaces and the adjoint)

Let $T : V \to V$ and let $S \subseteq V$ be a subspace.

(i) If S is T-invariant, then $S^\perp$ is T^*-invariant.

(ii) If S is invariant under both T and T^*, then, for the restricted inner product on S, $(T|_S)^* = T^*|_S$.

Proof. (i) Let $v \in S^{\perp}$ and $w \in S$. Then $T(w) \in S$, so $\langle T^*(v), w\rangle = \langle v, T(w)\rangle = 0$. Thus $T^*(v) \in S^{\perp}$.

(ii) The restriction of $\langle\cdot,\cdot\rangle$ to S is an inner product, and $T^*|_S$ maps S to S. For $v, w \in S$, $\langle v, T|_S(w)\rangle = \langle T^*|_S(v), w\rangle$, so $T^*|_S$ satisfies the defining property of the adjoint of $T|_S$; by uniqueness, $(T|_S)^* = T^*|_S$. □

9.4 Self-Adjoint and Normal Operators

We now study operators with special relationships to their adjoints. Throughout this section, V is a finite-dimensional inner product space.

9.4.1 Self-Adjoint Operators

The most important case is when $T = T^*$.

Definition 9.4.1 (Self-adjoint operator)

An operator $T : V \to V$ is ***self-adjoint*** (or ***Hermitian***) if $T^* = T$, i.e.,

$$\langle v, T(w)\rangle = \langle T(v), w\rangle$$

for all $v, w \in V$.

NOTE In an orthonormal basis, T is self-adjoint iff its matrix is symmetric (real case, $M = M^{t}$) or Hermitian (complex case, $M = M^{*}$), by Theorem 9.3.4. Self-adjoint operators need not be invertible: for example, the zero operator is self-adjoint.

Proposition 9.4.2 (Basic properties of self-adjoint operators)

Let $T : V \to V$ be self-adjoint.

(i) $\langle T(v), v\rangle$ is real for all $v \in V$.

(ii) Every eigenvalue of T is real.

(iii) Eigenvectors of T for distinct eigenvalues are orthogonal.

(iv) If $S \subseteq V$ is a T-invariant subspace, then $S^{\perp}$ is also T-invariant.

Proof. (i) $\langle T(v), v\rangle = \langle v, T(v)\rangle = \overline{\langle T(v), v\rangle}$.

(ii) If $T(v) = \lambda v$ with $v \neq 0$, then $\lambda\|v\|^2 = \langle v, T(v)\rangle = \langle T(v), v\rangle = \overline{\lambda}\|v\|^2$, so $\lambda = \overline{\lambda}$. (Over $\mathbb{R}$ there is nothing to prove; the content is in the complex case.)

(iii) If $T(v) = \lambda v$ and $T(w) = \mu w$ with $\lambda \neq \mu$, then λ is real by (ii), and

$$\mu\langle v, w\rangle = \langle v, T(w)\rangle = \langle T(v), w\rangle = \overline{\lambda}\langle v, w\rangle = \lambda\langle v, w\rangle.$$

So $(\mu - \lambda)\langle v, w\rangle = 0$, and $\langle v, w\rangle = 0$.

(iv) This is Lemma 9.3.7(i) with $T^* = T$. □

NOTE Property (iv) is what makes self-adjoint operators diagonalizable: once we find an eigenvector, its orthogonal complement is invariant, and we can proceed by induction.

9.4.2 The Spectral Theorem for Real Self-Adjoint Operators

Over $\mathbb{C}$, every operator has an eigenvector. Over $\mathbb{R}$, the main work is to show that a self-adjoint operator has one. The key is that quadratic polynomials without real roots cannot kill anything.

Lemma 9.4.3 (Positivity lemma)

Let T be self-adjoint on a finite-dimensional real inner product space. For any real $a > 0$, the operator $T^2 + a \cdot \mathrm{id}$ is invertible.

Proof. For $v \neq 0$, using $\langle T(T(v)), v\rangle = \langle T(v), T(v)\rangle$,

$$\langle (T^2 + a \cdot \mathrm{id})(v), v\rangle = \langle T^2(v), v\rangle + a\langle v, v\rangle = \|T(v)\|^2 + a\|v\|^2 > 0.$$

So $(T^2 + a \cdot \mathrm{id})(v) \neq 0$ for $v \neq 0$. Thus $T^2 + a \cdot \mathrm{id}$ is injective, hence invertible since V is finite-dimensional. □

Corollary 9.4.4 (Quadratics without real roots)

If $p(x) = x^2 + bx + c \in \mathbb{R}[x]$ has no real roots (i.e., $b^2 - 4c < 0$), and T is self-adjoint on a finite-dimensional real inner product space, then $p(T) = T^2 + bT + c \cdot \mathrm{id}$ is invertible.

Proof. Complete the square: $T^2 + bT + c \cdot \mathrm{id} = (T + \frac{b}{2}\mathrm{id})^2 + (c - \frac{b^2}{4})\mathrm{id}$.

Since $b^2 - 4c < 0$, we have $a = c - \frac{b^2}{4} > 0$. The operator $S = T + \frac{b}{2}\mathrm{id}$ is self-adjoint ($S^* = T^* + \frac{b}{2}\mathrm{id} = S$ by Proposition 9.3.3). By the lemma, $S^2 + a \cdot \mathrm{id} = p(T)$ is invertible. □

Theorem 9.4.5 (Spectral theorem (real self-adjoint operators))

Let $T : V \to V$ be a self-adjoint operator on a finite-dimensional real inner product space. Then V has an orthonormal basis consisting of eigenvectors

of T. In particular, T is diagonalizable, and its characteristic polynomial splits into linear factors over $\mathbb{R}$.

Proof. **Step 1: T has an eigenvector if $V \neq 0$.** Let $n = \dim V \geqslant 1$ and choose any nonzero $v \in V$. The $n+1$ vectors $v, T(v), \ldots, T^n(v)$ are linearly dependent, so there is a nonzero polynomial $p \in \mathbb{R}[x]$ with $p(T)(v) = 0$. It is not constant, since a nonzero multiple of v is nonzero.

Factor p over $\mathbb{R}$:

$$p(x) = a \prod_i (x - \lambda_i) \prod_j (x^2 + b_j x + c_j),$$

with $a \neq 0$, $\lambda_i \in \mathbb{R}$, and $b_j^2 - 4c_j < 0$. (Such a factorization exists by induction on $\deg p$. By the Fundamental Theorem of Algebra (Chapter 4), p has a root $\mu \in \mathbb{C}$. If μ is real, then $p = (x - \mu)q$ with $q \in \mathbb{R}[x]$ by the factor theorem (Lemma 4.3.14). Otherwise $p(\overline{\mu}) = \overline{p(\mu)} = 0$ because p has real coefficients, and since $\overline{\mu} \neq \mu$, applying the factor theorem twice in $\mathbb{C}[x]$ gives $p = mq$ with $m = (x - \mu)(x - \overline{\mu}) = x^2 - 2\operatorname{Re}(\mu)\, x + |\mu|^2$, a real quadratic with no real roots. Conjugating all coefficients gives $p = m\overline{q}$, so $m(q - \overline{q}) = 0$, and $q = \overline{q}$ has real coefficients because $\mathbb{C}[x]$ has no zero divisors (Lemma 4.3.3).) Since $q \mapsto q(T)$ is a ring homomorphism $\mathbb{R}[x] \to \operatorname{End}(V)$,

$$p(T) = a \prod_i (T - \lambda_i \mathrm{id}) \prod_j (T^2 + b_j T + c_j \mathrm{id}).$$

Each $T^2 + b_j T + c_j \mathrm{id}$ is invertible by the corollary. If every $T - \lambda_i \mathrm{id}$ were injective, then $p(T)$ would be a composite of injective operators, hence injective; but $p(T)(v) = 0$ with $v \neq 0$. So some $T - \lambda_i \mathrm{id}$ has a nonzero kernel, and T has an eigenvector with real eigenvalue λ_i.

Step 2: orthonormal eigenbasis, by induction on $\dim V$. If $\dim V = 0$, the empty list works. Otherwise, by Step 1 let v_1 be an eigenvector with $T(v_1) = \lambda_1 v_1$, normalized so $\|v_1\| = 1$. Let $S = \operatorname{span}(v_1)$. Since $T(S) \subseteq S$, the orthogonal complement $S^\perp$ is T-invariant (Proposition 9.4.2(iv)), and $\dim S^\perp = \dim V - 1$.

Since $S^\perp$ is invariant under $T = T^*$, the restriction $T|_{S^\perp}$ is self-adjoint for the restricted inner product (Lemma 9.3.7(ii)). By induction, $S^\perp$ has an orthonormal basis $\{v_2, \ldots, v_n\}$ of eigenvectors of T. Then $(v_1, v_2, \ldots, v_n)$ is an orthonormal list of length n, hence an orthonormal basis of V, consisting of eigenvectors.

Step 3: the characteristic polynomial. In this basis T has the diagonal matrix $\operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ with $\lambda_i \in \mathbb{R}$, so its characteristic polynomial is $\prod_i (x - \lambda_i)$ (Definition 7.4.7). □

NOTE The spectral theorem says that self-adjoint operators on real inner product spaces are "as nice as possible": they are diagonalizable, with real eigenvalues, and the eigenspaces are mutually orthogonal (Proposition 9.4.2(iii)). The matrix in an orthonormal eigenbasis is $\operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. In matrix terms: for every real symmetric matrix A there is a matrix P with $P^t P = I$ (its columns are the orthonormal eigenvectors) such that $P^{-1}AP = P^tAP$ is diagonal.

Skew-Adjoint Operators

The "opposite" of self-adjoint is skew-adjoint.

Definition 9.4.6 (Skew-adjoint operator)

An operator $T : V \to V$ is ***skew-adjoint*** (or ***anti-self-adjoint***) if $T^* = -T$, i.e.,

$$\langle v, T(w)\rangle = -\langle T(v), w\rangle$$

for all $v, w \in V$.

NOTE In an orthonormal basis, T is skew-adjoint iff its matrix M satisfies $M^t = -M$ (real case) or $M^* = -M$ (complex case). Such matrices are called ***skew-symmetric*** and ***skew-Hermitian***, respectively.

Proposition 9.4.7 (Properties of skew-adjoint operators)

Let $T : V \to V$ be skew-adjoint on a finite-dimensional real inner product space.

(i) $\langle T(v), v\rangle = 0$ for all $v \in V$.

(ii) The only possible eigenvalue of T is 0.

(iii) $\mathrm{id} + T$ is invertible.

Proof. (i) Taking $w = v$ in the definition gives $\langle v, T(v)\rangle = -\langle T(v), v\rangle$. Over $\mathbb{R}$, $\langle v, T(v)\rangle = \langle T(v), v\rangle$, so $2\langle T(v), v\rangle = 0$.

(ii) If $T(v) = \lambda v$ with $v \neq 0$, then $\lambda\|v\|^2 = \langle T(v), v\rangle = 0$ by (i), so $\lambda = 0$.

(iii) If $(\mathrm{id} + T)(v) = 0$, then $T(v) = -v$; by (ii), -1 is not an eigenvalue, so $v = 0$. Thus $\mathrm{id} + T$ is injective, hence invertible since V is finite-dimensional. □

NOTE Over $\mathbb{R}$, a skew-adjoint operator can have no eigenvalues at all: rotation by $\pi/2$ in $\mathbb{R}^2$, with matrix $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, is skew-adjoint. Its eigenvalues as a complex

matrix are $\pm i$; Corollary 9.4.13 below shows that the complex eigenvalues of a real skew-symmetric matrix are always purely imaginary.

9.4.3 Normal Operators and the Complex Spectral Theorem

Over $\mathbb{C}$, the spectral theorem extends to a larger class of operators: those that commute with their adjoint.

Definition 9.4.8 (Normal operator)

An operator $T : V \to V$ is ***normal*** if

$$TT^* = T^*T.$$

Example 9.4.9 (Classes of normal operators)

- Self-adjoint operators ($T^* = T$) are normal: $TT^* = T^2 = T^*T$.
- Skew-adjoint operators ($T^* = -T$) are normal: $TT^* = -T^2 = T^*T$.
- Operators with $T^* = T^{-1}$ are normal: $TT^* = \mathrm{id} = T^*T$. These are the orthogonal and unitary operators of Section 9.5.

Lemma 9.4.10 (Characterization of normality)

An operator T is normal if and only if $\|T(v)\| = \|T^*(v)\|$ for all $v \in V$.

Proof. By the defining property of the adjoint and $(T^*)^* = T$,

$$\|T(v)\|^2 = \langle T(v), T(v)\rangle = \langle T^*T(v), v\rangle,$$
$$\|T^*(v)\|^2 = \langle T^*(v), T^*(v)\rangle = \langle TT^*(v), v\rangle.$$

So $\|T(v)\| = \|T^*(v)\|$ for all v iff $\langle S(v), v\rangle = 0$ for all v, where $S = T^*T - TT^*$. If T is normal, $S = 0$. Conversely, S is self-adjoint (Proposition 9.3.3), and we claim that a self-adjoint S with $\langle S(v), v\rangle = 0$ for all v is zero. For any v, w,

$$\begin{aligned} 0 &= \langle S(v+w), v+w\rangle - \langle S(v), v\rangle - \langle S(w), w\rangle \\ &= \langle S(v), w\rangle + \langle S(w), v\rangle = \langle S(v), w\rangle + \overline{\langle S(v), w\rangle}, \end{aligned}$$

using $\langle S(w), v\rangle = \langle w, S(v)\rangle = \overline{\langle S(v), w\rangle}$. So $\operatorname{Re}\langle S(v), w\rangle = 0$. Over $\mathbb{R}$ this says $\langle S(v), w\rangle = 0$. Over $\mathbb{C}$, replacing w by iw gives $0 = \operatorname{Re}(i\langle S(v), w\rangle) = -\operatorname{Im}\langle S(v), w\rangle$. In both cases $\langle S(v), w\rangle = 0$ for all w; taking $w = S(v)$ gives $S(v) = 0$. □

Proposition 9.4.11 (Normal operators and eigenspaces)

Let T be normal.

(i) If $T(v) = \lambda v$, then $T^*(v) = \overline{\lambda} v$.

(ii) Eigenvectors for distinct eigenvalues are orthogonal.

(iii) $\ker(T - \lambda \mathrm{id}) = \ker(T^* - \overline{\lambda}\mathrm{id})$.

Proof. (i) The operator $T - \lambda\mathrm{id}$ is normal: its adjoint is $T^* - \overline{\lambda}\mathrm{id}$, and these two commute since T and T^* commute. By the lemma, $\|T^*(v) - \overline{\lambda}v\| = \|T(v) - \lambda v\| = 0$, so $T^*(v) = \overline{\lambda}v$.

(ii) If $T(v) = \lambda v$ and $T(w) = \mu w$ with $\lambda \neq \mu$, then by (i)

$$\mu\langle v, w\rangle = \langle v, T(w)\rangle = \langle T^*(v), w\rangle = \langle \overline{\lambda} v, w\rangle = \lambda\langle v, w\rangle.$$

So $(\mu - \lambda)\langle v, w\rangle = 0$, giving $v \perp w$.

(iii) The inclusion $\subseteq$ is (i). Since T^* is also normal, with $(T^*)^* = T$, applying (i) to T^* and the eigenvalue $\overline{\lambda}$ gives the reverse inclusion. □

Theorem 9.4.12 (Complex spectral theorem)

Let $T : V \to V$ be a normal operator on a finite-dimensional complex inner product space. Then V has an orthonormal basis consisting of eigenvectors of T. Moreover:

- if T is self-adjoint, its eigenvalues are real;
- if T is skew-adjoint, its eigenvalues are purely imaginary;
- if $T^* = T^{-1}$ (i.e., T is unitary, Section 9.5), its eigenvalues lie on the unit circle $|\lambda| = 1$.

Proof. We induct on $n = \dim V$; the case $n = 0$ is trivial. Let $n \geqslant 1$. Since $\mathbb{C}$ is algebraically closed, T has an eigenvector (Theorem 7.5.1); let v_1 be a unit eigenvector with $T(v_1) = \lambda v_1$.

By Proposition 9.4.11(i), $T^*(v_1) = \overline{\lambda} v_1$. So $S = \operatorname{span}(v_1)$ is invariant under both T and T^*. By Lemma 9.3.7(i), applied to T and to T^* (using $(T^*)^* = T$), the subspace $W = S^\perp$ is invariant under both T^* and T. By Lemma 9.3.7(ii), $(T|_W)^* = T^*|_W$, so

$$T|_W (T|_W)^* = (TT^*)|_W = (T^*T)|_W = (T|_W)^* T|_W,$$

and $T|_W$ is normal. Since $\dim W = n-1$, by induction W has an orthonormal basis $\{v_2, \ldots, v_n\}$ of eigenvectors of T. Then $(v_1, \ldots, v_n)$ is an orthonormal eigenbasis of V.

For the eigenvalues: if $T(v) = \lambda v$ with $v \neq 0$, then $T^*(v) = \overline{\lambda} v$. If $T^* = T$, this gives $\overline{\lambda} = \lambda$. If $T^* = -T$, it gives $\overline{\lambda} = -\lambda$. If $T^* = T^{-1}$, then $v = T^*(T(v)) = \overline{\lambda}\lambda v$, so $|\lambda|^2 = 1$. □

NOTE The complex spectral theorem is stronger than the real one: every normal operator (not just every self-adjoint one) is diagonalizable in an orthonormal basis. The price is working over $\mathbb{C}$ with Hermitian inner products. Normality is also necessary: if T has an orthonormal eigenbasis, its matrix D in that basis is diagonal, T^* has the diagonal matrix $D^* = \overline{D}$, and diagonal matrices commute, so T is normal.

Corollary 9.4.13 (Eigenvalues of real symmetric and skew-symmetric matrices)

Let A be a real $n \times n$ matrix, and consider its complex eigenvalues (the roots in $\mathbb{C}$ of its characteristic polynomial).

(i) If $A^t = A$, all complex eigenvalues of A are real.

(ii) If $A^t = -A$, all complex eigenvalues of A are purely imaginary.

(iii) If $A^t A = I$, all complex eigenvalues of A have absolute value 1.

Proof. Let $T : \mathbb{C}^n \to \mathbb{C}^n$, $T(z) = Az$, with the standard Hermitian inner product. The standard basis is orthonormal, so T^* has matrix $A^* = A^t$ (the entries of A are real). Thus T is self-adjoint in case (i), skew-adjoint in case (ii), and satisfies $T^* = T^{-1}$ in case (iii). The roots of the characteristic polynomial of A are the eigenvalues of T, and the theorem applies. □

9.5 Orthogonal and Unitary Operators

9.5.1 Definitions and Characterizations

Orthogonal (real) and unitary (complex) operators are the isometries of inner product spaces: they preserve lengths and angles.

Definition 9.5.1 (Orthogonal and unitary operators)

Let V be an inner product space. An operator $T : V \to V$ is ***orthogonal*** (if V is real) or ***unitary*** (if V is complex) if

$$\langle T(u), T(v) \rangle = \langle u, v \rangle$$

for all $u, v \in V$.

NOTE The definitions are parallel; some authors use "orthogonal" in both cases. Note that an orthogonal operator is not the same thing as an orthogonal projection.

Theorem 9.5.2 (Characterizations of orthogonal and unitary operators)

For an operator $T : V \to V$ on a finite-dimensional inner product space, the following are equivalent:

(i) T is orthogonal (resp. unitary).
(ii) $T^*T = \mathrm{id}$, i.e., T is invertible and $T^* = T^{-1}$.
(iii) $TT^* = \mathrm{id}$.
(iv) $\|T(v)\| = \|v\|$ for all $v \in V$.
(v) T maps some orthonormal basis to an orthonormal basis.
(vi) T maps every orthonormal basis to an orthonormal basis.

Proof. (i) $\Rightarrow$ (ii): $\langle T^*T(u), v\rangle = \langle T(u), T(v)\rangle = \langle u, v\rangle$ for all u, v, so $T^*T(u) - u \in V^\perp = \{0\}$.

(ii) $\Leftrightarrow$ (iii): On a finite-dimensional space, a one-sided inverse of an operator is two-sided: $T^*T = \mathrm{id}$ makes T injective, hence invertible with $T^{-1} = T^*$, so $TT^* = \mathrm{id}$; and $TT^* = \mathrm{id}$ makes T surjective, hence invertible with $T^{-1} = T^*$, so $T^*T = \mathrm{id}$.

(ii) $\Rightarrow$ (iv): $\|T(v)\|^2 = \langle T(v), T(v)\rangle = \langle T^*T(v), v\rangle = \langle v, v\rangle = \|v\|^2$.

(iv) $\Rightarrow$ (i): The inner product can be recovered from norms by the ***polarization identities***. Over $\mathbb{R}$, $\|u \pm v\|^2 = \|u\|^2 \pm 2\langle u, v\rangle + \|v\|^2$, so

$$\langle u, v\rangle = \tfrac{1}{4}\left(\|u + v\|^2 - \|u - v\|^2\right).$$

Over $\mathbb{C}$, for $m = 0, 1, 2, 3$ we have $\|u + i^m v\|^2 = \|u\|^2 + \|v\|^2 + i^m\langle u, v\rangle + (-i)^m\langle v, u\rangle$. Multiplying by $(-i)^m$ and summing over m, and using $\sum_{m=0}^{3}(-i)^m = 0$ and $\sum_{m=0}^{3}(-1)^m = 0$, gives

$$\langle u, v\rangle = \frac{1}{4}\sum_{m=0}^{3}(-i)^m \|u + i^m v\|^2.$$

In both cases the right side is unchanged when u, v are replaced by $T(u), T(v)$, since $\|T(u) + i^m T(v)\| = \|T(u + i^m v)\| = \|u + i^m v\|$. So $\langle T(u), T(v)\rangle = \langle u, v\rangle$.

(i) $\Rightarrow$ (vi): If $\{e_1, \ldots, e_n\}$ is an orthonormal basis, then $\langle T(e_i), T(e_j)\rangle = \langle e_i, e_j\rangle = \delta_{ij}$, so $(T(e_1), \ldots, T(e_n))$ is an orthonormal list of length n, hence an orthonormal basis.

(vi) $\Rightarrow$ (v): Orthonormal bases exist (Theorem 9.2.2).

(v) $\Rightarrow$ (ii): Let $\{e_i\}$ be an orthonormal basis with $\{T(e_i)\}$ orthonormal. Then

$$\langle T^*T(e_i), e_j\rangle = \langle T(e_i), T(e_j)\rangle = \delta_{ij} = \langle e_i, e_j\rangle$$

for all i, j. Since the e_j span V, $T^*T(e_i) - e_i \in V^\perp = \{0\}$ for each i, so $T^*T = \mathrm{id}$. □

Corollary 9.5.3 (Orthogonal and unitary operators form a group)

The orthogonal (resp. unitary) operators on a finite-dimensional inner product space V form a subgroup of $\mathrm{GL}(V)$, denoted $\mathrm{O}(V)$ (resp. $\mathrm{U}(V)$). If T is orthogonal or unitary, then so is $T^{-1} = T^*$.

Proof. By the theorem, such operators are invertible. The identity preserves the inner product. If S and T do, then $\langle ST(u), ST(v)\rangle = \langle T(u), T(v)\rangle = \langle u, v\rangle$. And $\langle T^{-1}(u), T^{-1}(v)\rangle = \langle T(T^{-1}(u)), T(T^{-1}(v))\rangle = \langle u, v\rangle$. □

Proposition 9.5.4 (Eigenvalues and invariant subspaces)

Let T be an orthogonal or unitary operator on a finite-dimensional inner product space V.

(i) Every eigenvalue λ of T satisfies $|\lambda| = 1$. In the real case, $\lambda = \pm 1$.

(ii) If $S \subseteq V$ is T-invariant, then $T(S) = S$, and $S^\perp$ is also T-invariant.

Proof. (i) If $T(v) = \lambda v$ with $v \neq 0$, then $\|v\| = \|T(v)\| = |\lambda|\|v\|$, so $|\lambda| = 1$.

(ii) $T|_S : S \to S$ is injective and S is finite-dimensional, so $T(S) = S$. Hence $T^{-1}(S) = S$, i.e., S is invariant under $T^{-1} = T^*$. By Lemma 9.3.7(i) applied to T^* (whose adjoint is T), $S^\perp$ is T-invariant. □

Matrix Characterization

Proposition 9.5.5 (Orthogonal and unitary matrices)

Let $\{e_1, \ldots, e_n\}$ be an orthonormal basis of V, and let M be the matrix of T in this basis.

(i) (Real case) T is orthogonal iff $M^tM = I$; such M is called an ***orthogonal matrix***.

(ii) (Complex case) T is unitary iff $M^*M = I$; such M is called a ***unitary matrix***.

Equivalently: the columns of M form an orthonormal basis of k^n with its standard inner product.

Proof. In an orthonormal basis, $\mathcal{M}(T^*) = M^*$ (Theorem 9.3.4), so $T^*T = \mathrm{id}$ iff $M^*M = I$.

Write $c_1, \dots, c_n$ for the columns of M. The (i,j)-entry of M^*M is $\sum_k \overline{M_{ki}}\, M_{kj} = \langle c_i, c_j \rangle$, the standard inner product. So $M^*M = I$ iff the columns are orthonormal. □

Corollary 9.5.6 (Rows of orthogonal and unitary matrices)

If M is orthogonal (resp. unitary), then the rows of M also form an orthonormal basis of k^n.

Proof. $M^*M = I$ implies that M is invertible with $M^{-1} = M^*$, so $MM^* = I$. Writing $r_1, \dots, r_n$ for the rows of M (as vectors in k^n), the (i,j)-entry of MM^* is $\sum_k M_{ik}\, \overline{M_{jk}} = \langle r_j, r_i \rangle$. So the rows are orthonormal. □

Proposition 9.5.7 (Determinant of orthogonal and unitary operators)

If T is orthogonal or unitary, then $|\det(T)| = 1$. For orthogonal T, $\det(T) = \pm 1$.

Proof. Let M be the matrix of T in an orthonormal basis. Then T^* has matrix M^* (Theorem 9.3.4), and $\det(M^*) = \overline{\det M}$ (Example 5.4.25; over $\mathbb{R}$ this is $\det(M^t) = \det M$, Proposition 5.4.6). Hence, by multiplicativity (Theorem 5.4.7),

$$1 = \det(\mathrm{id}) = \det(T^*T) = \det(T^*)\det(T) = \overline{\det(T)}\,\det(T) = |\det(T)|^2.$$

Over $\mathbb{R}$, $\det(T)$ is a real number of absolute value 1, so $\det(T) = \pm 1$. □

Skew-adjoint operators give a supply of orthogonal ones.

Proposition 9.5.8 (Cayley transform)

Let T be skew-adjoint on a finite-dimensional real inner product space. Then $\mathrm{id} + T$ and $\mathrm{id} - T$ are invertible, and

$$S = (\mathrm{id} - T)(\mathrm{id} + T)^{-1}$$

is orthogonal.

Proof. $\mathrm{id} + T$ is invertible by Proposition 9.4.7(iii), and so is $\mathrm{id} - T = \mathrm{id} + (-T)$, since $-T$ is also skew-adjoint. Since $T^* = -T$, we have $(\mathrm{id} + T)^* = \mathrm{id} - T$ and $(\mathrm{id} - T)^* = \mathrm{id} + T$. The operators $\mathrm{id} - T$ and $\mathrm{id} + T$ commute (both are polynomials in T). Using Proposition 9.3.3,

$$S^* = \big((\mathrm{id} + T)^{-1}\big)^*(\mathrm{id} - T)^* = \big((\mathrm{id} + T)^*\big)^{-1}(\mathrm{id} + T) = (\mathrm{id} - T)^{-1}(\mathrm{id} + T).$$

Hence

$$\begin{aligned} S^*S &= (\mathrm{id}-T)^{-1}(\mathrm{id}+T)(\mathrm{id}-T)(\mathrm{id}+T)^{-1} \\ &= (\mathrm{id}-T)^{-1}(\mathrm{id}-T)(\mathrm{id}+T)(\mathrm{id}+T)^{-1} = \mathrm{id}, \end{aligned}$$

and S is orthogonal by Theorem 9.5.2. □

9.5.2 Reflections

A particularly simple class of orthogonal operators are the reflections across hyperplanes.

Definition 9.5.9 (Reflection)

Let V be a finite-dimensional real inner product space. A ***reflection*** is an orthogonal operator $P : V \to V$ such that $P^2 = \mathrm{id}$ and $\dim \ker(P + \mathrm{id}) = 1$.

NOTE The condition $P^2 = \mathrm{id}$ says P is an involution. Since P is orthogonal, it is also self-adjoint ($P^* = P^{-1} = P$). The only possible eigenvalues of P are ± 1 (if $P(v) = \lambda v$ with $v \neq 0$, then $v = P^2(v) = \lambda^2 v$), and $\dim \ker(P + \mathrm{id}) = 1$ says the -1-eigenspace is a line.

Proposition 9.5.10 (Formula for reflections)

For a unit vector $v \in V$, the map

$$P_v(x) = x - 2\langle x, v\rangle v$$

is a reflection, with $\ker(P_v + \mathrm{id}) = \mathrm{span}(v)$, and P_v fixes every vector of the hyperplane $v^\perp$. Conversely, every reflection is P_v for a unit vector v spanning $\ker(P + \mathrm{id})$. Geometrically, P_v is the reflection across the hyperplane $v^\perp$.

Proof. **Linear:** Clear from the formula.

Orthogonal:

$$\begin{aligned} \langle P_v(x), P_v(y)\rangle &= \langle x - 2\langle x, v\rangle v,\ y - 2\langle y, v\rangle v\rangle \\ &= \langle x, y\rangle - 2\langle x, v\rangle\langle v, y\rangle - 2\langle y, v\rangle\langle x, v\rangle + 4\langle x, v\rangle\langle y, v\rangle\|v\|^2 \\ &= \langle x, y\rangle - 4\langle x, v\rangle\langle y, v\rangle + 4\langle x, v\rangle\langle y, v\rangle = \langle x, y\rangle. \end{aligned}$$

Involution: $\langle P_v(x), v\rangle = \langle x, v\rangle - 2\langle x, v\rangle\|v\|^2 = -\langle x, v\rangle$, so

$$P_v(P_v(x)) = P_v(x) + 2\langle x, v\rangle v = x.$$

Eigenspaces: $P_v(x) = -x$ iff $x = \langle x, v\rangle v$, iff $x \in \operatorname{span}(v)$ (if $x = cv$, then $\langle x, v\rangle v = cv$). So $\ker(P_v + \mathrm{id}) = \operatorname{span}(v)$ is a line. If $w \perp v$, then $P_v(w) = w$.

Converse: Let P be a reflection and v a unit vector spanning $\ker(P + \mathrm{id})$, so $P(v) = -v$. If $x \in v^\perp$, then $\langle P(x), v\rangle = -\langle P(x), P(v)\rangle = -\langle x, v\rangle = 0$, so $P(x) \in v^\perp$. The vector $w = x - P(x)$ then lies in $v^\perp$ and satisfies $P(w) = P(x) - x = -w$, so $w \in \operatorname{span}(v) \cap v^\perp = \{0\}$. Hence P fixes $v^\perp$ pointwise. Writing any x as $x = \langle x, v\rangle v + (x - \langle x, v\rangle v)$ with the second term in $v^\perp$,

$$P(x) = -\langle x, v\rangle v + (x - \langle x, v\rangle v) = x - 2\langle x, v\rangle v = P_v(x). \qquad \square$$

Proposition 9.5.11 (Reflections map equal-length vectors)

Let $x, y \in V$ with $\|x\| = \|y\|$.

(i) If $x \neq y$, the reflection P_v with $v = (x - y)/\|x - y\|$ satisfies $P_v(x) = y$.

(ii) If $x = y$ and $\dim V \geqslant 2$, there is a reflection P with $P(x) = y$.

Proof. (i) Since $\|x\|^2 = \|y\|^2$,

$$\langle x, x - y\rangle = \|x\|^2 - \langle x, y\rangle = \tfrac{1}{2}\big(\|x\|^2 + \|y\|^2 - 2\langle x, y\rangle\big) = \tfrac{1}{2}\|x - y\|^2.$$

Therefore

$$P_v(x) = x - \frac{2\langle x, x - y\rangle}{\|x - y\|^2}(x - y) = x - (x - y) = y.$$

(ii) Since $\dim x^\perp \geqslant \dim V - 1 \geqslant 1$, we can choose a unit vector $v \perp x$; then $P_v(x) = x - 2\langle x, v\rangle v = x$. $\square$

NOTE The hypothesis in (ii) is needed: if $\dim V = 1$, the only reflection is $-\mathrm{id}$, which fixes only 0.

Theorem 9.5.12 (Orthogonal operators as products of reflections)

Let V be a real inner product space of dimension n. Every orthogonal operator on V is a product of at most n reflections. (The identity is the empty product.)

Proof. Induction on n. For $n = 0$, the only operator is the identity, the empty product.

For $n \geqslant 1$, let T be orthogonal and pick a unit vector e_1. Then $\|T(e_1)\| = \|e_1\| = 1$. If $T(e_1) = e_1$, let $T_1 = T$. Otherwise, by Proposition 9.5.11(i) there is a reflection P_1 with $P_1(T(e_1)) = e_1$; let $T_1 = P_1 T$. In both cases T_1 is orthogonal and $T_1(e_1) = e_1$.

Since $\operatorname{span}(e_1)$ is T_1-invariant, so is $W = e_1^{\perp}$ (Proposition 9.5.4(ii)), and $T_1|_W$ is an orthogonal operator on the $(n-1)$-dimensional space W. By induction, $T_1|_W = Q_1 \cdots Q_m$ for some reflections Q_i of W with $m \leqslant n-1$.

Extend each Q_i to $V = \operatorname{span}(e_1) \oplus W$ by setting $\tilde{Q}_i(ae_1 + w) = ae_1 + Q_i(w)$. Then $\tilde{Q}_i$ is orthogonal (because $e_1 \perp W$ and Q_i is orthogonal on W) and $\tilde{Q}_i^2 = \mathrm{id}$. Moreover, $\tilde{Q}_i(ae_1 + w) = -(ae_1 + w)$ iff $a = 0$ and $Q_i(w) = -w$, so $\ker(\tilde{Q}_i + \mathrm{id}) = \ker(Q_i + \mathrm{id})$ is a line, and $\tilde{Q}_i$ is a reflection.

Now $T_1 = \tilde{Q}_1 \cdots \tilde{Q}_m$, since both sides fix e_1 and agree on W. If $T_1 = T$, we are done. If $T_1 = P_1 T$, then since $P_1^2 = \mathrm{id}$,

$$T = P_1 \tilde{Q}_1 \cdots \tilde{Q}_m,$$

a product of at most $1 + (n-1) = n$ reflections. □

9.5.3 The Classical Matrix Groups

The orthogonal and unitary operators, written as matrices in orthonormal bases, form important matrix groups.

Definition 9.5.13 (Orthogonal and special orthogonal groups)

- $O(n) = \{A \in \mathrm{GL}_n(\mathbb{R}) : A^{t}A = I\}$, the ***orthogonal group***.
- $SO(n) = \{A \in O(n) : \det(A) = 1\}$, the ***special orthogonal group***.

Definition 9.5.14 (Unitary and special unitary groups)

- $U(n) = \{A \in \mathrm{GL}_n(\mathbb{C}) : A^{*}A = I\}$, the ***unitary group***.
- $SU(n) = \{A \in U(n) : \det(A) = 1\}$, the ***special unitary group***.

NOTE By Proposition 9.5.5, choosing an orthonormal basis of an n-dimensional real inner product space V identifies $O(V)$ with $O(n)$, and similarly $U(V) \cong U(n)$ in the complex case. We write $SO(V)$ for the orthogonal operators of determinant 1.

Proposition 9.5.15 (Index of $SO(n)$ in $O(n)$)

For $n \geqslant 1$, $SO(n)$ is a normal subgroup of $O(n)$ of index 2. We have an exact sequence

$$1 \to SO(n) \to O(n) \xrightarrow{\det} \{\pm 1\} \to 1.$$

Proof. The determinant of an orthogonal matrix is ± 1, and det is multiplicative, so $\det : \mathrm{O}(n) \to \{\pm 1\}$ is a homomorphism. It is surjective because $\det(I) = 1$ and the reflection matrix $\mathrm{diag}(-1, 1, \ldots, 1) \in \mathrm{O}(n)$ has determinant -1. Its kernel is $\mathrm{SO}(n)$, which is therefore normal, and $\mathrm{O}(n)/\mathrm{SO}(n) \cong \{\pm 1\}$ by the First Isomorphism Theorem (Theorem 3.3.3). □

NOTE Elements of $\mathrm{SO}(n)$ are called ***rotations*** or ***proper orthogonal transformations***; they are the orthogonal transformations that preserve orientation, in the sense that they have positive determinant. Elements of $\mathrm{O}(n) \setminus \mathrm{SO}(n)$, such as reflections, reverse orientation.

In dimension 2 everything can be made explicit. Write

$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \qquad S_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}.$$

Proposition 9.5.16 (The group O(2))

Every element of $\mathrm{O}(2)$ is either a rotation R_θ (determinant 1) or a matrix S_θ (determinant -1), for a unique $\theta \in [0, 2\pi)$. The matrix S_θ is the reflection across the line spanned by $(\cos\frac{\theta}{2}, \sin\frac{\theta}{2})$: it fixes this vector and negates the orthogonal vector $(-\sin\frac{\theta}{2}, \cos\frac{\theta}{2})$. If $\theta \notin \{0, \pi\}$, the rotation R_θ has no real eigenvalues.

Proof. Let $A \in \mathrm{O}(2)$. By Proposition 9.5.5, its columns c_1, c_2 are orthonormal in $\mathbb{R}^2$. Since c_1 is a unit vector, $c_1 = (\cos\theta, \sin\theta)$ for a unique $\theta \in [0, 2\pi)$. The vector c_2 is a unit vector in the line $c_1^\perp = \mathrm{span}((-\sin\theta, \cos\theta))$, so $c_2 = \pm(-\sin\theta, \cos\theta)$. The sign $+$ gives R_θ, with $\det R_\theta = \cos^2\theta + \sin^2\theta = 1$, and the sign $-$ gives S_θ, with $\det S_\theta = -1$.

Let $\varphi = \theta/2$. Using $\cos\theta\cos\varphi + \sin\theta\sin\varphi = \cos(\theta - \varphi)$ and $\sin\theta\cos\varphi - \cos\theta\sin\varphi = \sin(\theta - \varphi)$,

$$S_\theta \begin{pmatrix} \cos\varphi \\ \sin\varphi \end{pmatrix} = \begin{pmatrix} \cos\varphi \\ \sin\varphi \end{pmatrix}, \qquad S_\theta \begin{pmatrix} -\sin\varphi \\ \cos\varphi \end{pmatrix} = \begin{pmatrix} \sin\varphi \\ -\cos\varphi \end{pmatrix} = -\begin{pmatrix} -\sin\varphi \\ \cos\varphi \end{pmatrix}.$$

These two vectors form an orthonormal basis, so S_θ is the reflection P_ν with $\nu = (-\sin\varphi, \cos\varphi)$ (Proposition 9.5.10).

Finally, the characteristic polynomial of R_θ is $x^2 - 2\cos\theta\, x + 1$, whose discriminant $4\cos^2\theta - 4$ is negative unless $\cos\theta = \pm 1$, i.e., unless $\theta \in \{0, \pi\}$. □

Proposition 9.5.17 (SO(2) is the circle group)

$\mathrm{SO}(2) = \{R_\theta : \theta \in \mathbb{R}\}$, and $R_\theta \mapsto e^{i\theta}$ is a well-defined group isomorphism from $\mathrm{SO}(2)$ to the circle group $S^1 = \{z \in \mathbb{C} : |z| = 1\}$.

Proof. By Proposition 9.5.16, the elements of $\mathrm{SO}(2)$ are exactly the R_θ. Since R_θ is determined by its first column $(\cos\theta, \sin\theta)$, we have $R_\theta = R_\varphi$ iff $e^{i\theta} = e^{i\varphi}$; so the map is well defined and injective. It is surjective because every complex number of absolute value 1 is $e^{i\theta}$ for some θ. By the addition formulas for sine and cosine,

$$R_\theta R_\varphi = \begin{pmatrix} \cos\theta\cos\varphi - \sin\theta\sin\varphi & -(\sin\theta\cos\varphi + \cos\theta\sin\varphi) \\ \sin\theta\cos\varphi + \cos\theta\sin\varphi & \cos\theta\cos\varphi - \sin\theta\sin\varphi \end{pmatrix} = R_{\theta+\varphi},$$

which corresponds to $e^{i\theta}e^{i\varphi} = e^{i(\theta+\varphi)}$. So the map is a homomorphism. □

Example 9.5.18 (Low-dimensional cases)

- $\mathrm{O}(1) = \{\pm 1\} \cong \mathbb{Z}/2$, and $\mathrm{SO}(1) = \{1\}$.
- $\mathrm{SO}(2) \cong S^1$ (Proposition 9.5.17), while $\mathrm{O}(2)$ consists of the rotations R_θ and the reflections S_θ.
- $\mathrm{U}(1) = \{z \in \mathbb{C} : |z| = 1\} = S^1$, acting on $\mathbb{C}$ by multiplication.

The structure of a single orthogonal operator is now a consequence of the existence of small invariant subspaces.

Theorem 9.5.19 (Structure of orthogonal operators)

Let T be an orthogonal operator on a finite-dimensional real inner product space V. Then V is an orthogonal direct sum (the summands are pairwise orthogonal)

$$V = V_+ \oplus V_- \oplus W_1 \oplus \cdots \oplus W_k,$$

where $V_+ = \ker(T - \mathrm{id})$, $V_- = \ker(T + \mathrm{id})$, and each W_j is a 2-dimensional T-invariant subspace with an orthonormal basis in which $T|_{W_j}$ has matrix R_{θ_j} for some $\theta_j \in (0, \pi)$.

Proof. **Step 1: a decomposition into small blocks.** We show by induction on $\dim V$ that V is an orthogonal direct sum of T-invariant subspaces, each of which is either a line on which $T = \pm\mathrm{id}$ or a plane with an orthonormal basis in which T has matrix R_θ with $\theta \in (0, \pi)$. If $V = 0$ there is nothing to prove.

If $V \neq 0$, then T has an invariant subspace U of dimension 1 or 2 (Theorem 7.10.7). The restriction $T|_U$ is orthogonal for the restricted inner product, $V = U \oplus U^\perp$ (Theorem 9.1.31), and $U^\perp$ is T-invariant by Proposition 9.5.4(ii),

with $T|_{U^\perp}$ orthogonal. By induction, $U^\perp$ decomposes as required, so it remains to decompose U.

If $\dim U = 1$, then T acts on U by an eigenvalue, which is ± 1 by Proposition 9.5.4(i). If $\dim U = 2$, choose an orthonormal basis (u_1, u_2) of U; the matrix of $T|_U$ lies in $O(2)$, so by Proposition 9.5.16 it is S_θ or R_θ. If it is S_θ, then U is the orthogonal direct sum of two lines on which T acts by $+1$ and -1. If it is R_θ with $\theta \in \{0, \pi\}$, then $T|_U = \pm\mathrm{id}$ and U is the orthogonal direct sum of two lines. If it is R_θ with $\theta \in (\pi, 2\pi)$, replacing u_2 by $-u_2$ changes the matrix to $R_{-\theta} = R_{2\pi-\theta}$ with $2\pi - \theta \in (0, \pi)$.

Step 2: identifying $V_\pm$. Let V'_+ (resp. V'_-) be the sum of the lines in Step 1 on which $T = \mathrm{id}$ (resp. $T = -\mathrm{id}$), and $W_1, \dots, W_k$ the planes. Then $V'_+ \subseteq \ker(T - \mathrm{id})$. Conversely, $T - \mathrm{id}$ is injective on each of the other blocks: it is $-2\,\mathrm{id}$ on V'_-, and on W_j its matrix $R_{\theta_j} - I$ has determinant $(\cos\theta_j - 1)^2 + \sin^2\theta_j = 2 - 2\cos\theta_j \neq 0$. Since $T - \mathrm{id}$ preserves each block, a vector $v = v_+ + v_- + w_1 + \cdots + w_k$ with $(T - \mathrm{id})(v) = 0$ has $(T - \mathrm{id})$ of each component equal to 0, so $v = v_+ \in V'_+$. Thus $V'_+ = \ker(T - \mathrm{id})$, and similarly $V'_- = \ker(T + \mathrm{id})$, using $\det(R_{\theta_j} + I) = 2 + 2\cos\theta_j \neq 0$. □

NOTE In matrix form: in an orthonormal basis adapted to this decomposition, T has the block-diagonal matrix

$$\begin{pmatrix} I_a & & & & \\ & -I_b & & & \\ & & R_{\theta_1} & & \\ & & & \ddots & \\ & & & & R_{\theta_k} \end{pmatrix},$$

the ***real canonical form*** of an orthogonal matrix. This is a very good way to think about a single orthogonal transformation, built from rotations of individual planes and reflections; it says little about the composition of two orthogonal transformations, whose invariant subspaces are in general unrelated.

Exercise 9.5.20 (Inner product spaces)

(1) Let $V \subset \mathbb{R}[x]$ be the space of polynomials of degree at most 2, with inner product $\langle f, g\rangle = \int_0^1 f(x)g(x)\,dx$. Apply Gram–Schmidt to the basis $\{1, x, x^2\}$ to find an orthonormal basis.

(2) Let $T : V \to V$ be self-adjoint on a finite-dimensional real inner product space, and suppose $v \in V$ with $\|v\| = 1$ maximizes the Rayleigh

quotient $R(w) = \langle T(w), w\rangle / \|w\|^2$ among unit vectors. Prove that v is an eigenvector of T. (Do not use the spectral theorem.)

(3) Let $T : V \to V$ be an invertible linear operator on a finite-dimensional real inner product space. Show that T maps the unit sphere $S = \{v : \|v\| = 1\}$ to an ellipsoid. (Hint: consider $(T^{-1})^* T^{-1}$.)

(4) For an invertible operator T on a finite-dimensional real inner product space, define the ***dilatation*** $K(T) = \sup_{\|v\|=1} \|T(v)\| \,/\, \inf_{\|v\|=1} \|T(v)\|$. Show that $K(T)^2$ equals the ratio of the largest to the smallest eigenvalue of T^*T.

(5) Let A be the matrix of an inner product on $\mathbb{R}^n$ with respect to the standard basis, so that $\langle x, y\rangle = x^t A y$. For $n = 2$, prove that an entry of A of largest absolute value occurs on the diagonal.

Solution. **(1)** Start with $w_1 = 1$, $w_2 = x$, $w_3 = x^2$.
$\langle w_1, w_1\rangle = \int_0^1 1\,dx = 1$, so $v_1 = 1$.
$\langle v_1, w_2\rangle = \int_0^1 x\,dx = \frac{1}{2}$, so $u_2 = x - \frac{1}{2}\cdot 1 = x - \frac{1}{2}$, and $\|u_2\|^2 = \int_0^1 (x-\frac{1}{2})^2 dx = \int_0^1 (x^2 - x + \frac{1}{4})dx = \frac{1}{3} - \frac{1}{2} + \frac{1}{4} = \frac{1}{12}$. So $v_2 = \sqrt{12}(x - \frac{1}{2}) = 2\sqrt{3}(x - \frac{1}{2})$.
$\langle v_1, w_3\rangle = \int_0^1 x^2 dx = \frac{1}{3}$ and $\langle v_2, w_3\rangle = 2\sqrt{3}\int_0^1 x^2(x-\frac{1}{2})dx = 2\sqrt{3}(\frac{1}{4} - \frac{1}{6}) = \frac{\sqrt{3}}{6}$, so

$$u_3 = x^2 - \frac{1}{3} - \frac{\sqrt{3}}{6}\cdot 2\sqrt{3}\left(x - \frac{1}{2}\right) = x^2 - \frac{1}{3} - \left(x - \frac{1}{2}\right) = x^2 - x + \frac{1}{6},$$

and $\|u_3\|^2 = \int_0^1 (x^2 - x + \frac{1}{6})^2 dx = \frac{1}{180}$. So $v_3 = \sqrt{180}(x^2 - x + \frac{1}{6}) = 6\sqrt{5}(x^2 - x + \frac{1}{6})$.
The orthonormal basis is $\{1,\ 2\sqrt{3}(x - \frac{1}{2}),\ 6\sqrt{5}(x^2 - x + \frac{1}{6})\}$.
(2) Since $R(cw) = R(w)$ for $c \neq 0$, v maximizes R over all nonzero vectors. Let $u \perp v$ and $t \in \mathbb{R}$, and put $w_t = v + tu$ (nonzero, since $\|w_t\|^2 = 1 + t^2\|u\|^2$). Since T is self-adjoint and the space is real, $\langle T(u), v\rangle = \langle u, T(v)\rangle = \langle T(v), u\rangle$, so

$$\langle T(w_t), w_t\rangle = \langle T(v), v\rangle + 2t\langle T(v), u\rangle + t^2\langle T(u), u\rangle.$$

Thus

$$R(w_t) = \frac{\langle T(v), v\rangle + 2t\langle T(v), u\rangle + t^2\langle T(u), u\rangle}{1 + t^2\|u\|^2}.$$

This differentiable function of t has a maximum at $t = 0$, so its derivative there vanishes. The derivative at $t = 0$ is $2\langle T(v), u\rangle$. So $\langle T(v), u\rangle = 0$ for all $u \perp v$.
Thus $T(v) \in (v^\perp)^\perp = \operatorname{span}(v)$ (Corollary 9.1.32), i.e., $T(v) = \lambda v$ for some λ.

(3) We have $T(S) = \{w : \|T^{-1}(w)\| = 1\}$. Now

$$\|T^{-1}(w)\|^2 = \langle T^{-1}(w),\, T^{-1}(w)\rangle = \langle (T^{-1})^* T^{-1}(w),\, w\rangle = \langle Q(w),\, w\rangle,$$

where $Q = (T^{-1})^* T^{-1}$ is self-adjoint (Proposition 9.3.3) and satisfies $\langle Q(w), w\rangle = \|T^{-1}(w)\|^2 > 0$ for $w \neq 0$.
By the spectral theorem, there exists an orthonormal eigenbasis $\{u_1, \ldots, u_n\}$ with $Q(u_i) = \mu_i u_i$, where $\mu_i = \langle Q(u_i), u_i\rangle > 0$.
For $w = \sum x_i u_i$: $\langle Q(w), w\rangle = \sum \mu_i x_i^2$.
So $T(S) = \{w : \sum \mu_i x_i^2 = 1\}$. Setting $a_i = 1/\sqrt{\mu_i}$, this is $\sum x_i^2/a_i^2 = 1$, an ellipsoid with semi-axes a_i along u_i.
(4) We have $\|T(v)\|^2 = \langle T(v), T(v)\rangle = \langle T^*T(v), v\rangle$. The operator T^*T is self-adjoint, and $\langle T^*T(v), v\rangle = \|T(v)\|^2 > 0$ for $v \neq 0$ (since T is invertible).
By the spectral theorem, let $\{e_i\}$ be an orthonormal eigenbasis with $T^*T(e_i) = \lambda_i e_i$; then $\lambda_i = \|T(e_i)\|^2 > 0$. For $v = \sum a_i e_i$ with $\|v\| = 1$ (i.e., $\sum a_i^2 = 1$):

$$\|T(v)\|^2 = \langle T^*T(v), v\rangle = \sum \lambda_i a_i^2.$$

This weighted average of the λ_i is at most $\lambda_{\max}$ and at least $\lambda_{\min}$, with equality when v is a corresponding eigenvector. So $\sup_{\|v\|=1} \|T(v)\|^2 = \lambda_{\max}$ and $\inf_{\|v\|=1} \|T(v)\|^2 = \lambda_{\min}$.
Thus $K(T)^2 = \lambda_{\max}/\lambda_{\min}$.
(5) For $n = 2$, let $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$. Positive-definiteness gives $a = \langle e_1, e_1\rangle > 0$ and $c = \langle e_2, e_2\rangle > 0$, and, evaluating at $x = (-b, a)$,

$$\langle x, x\rangle = ab^2 - 2ab^2 + ca^2 = a(ac - b^2) > 0,$$

so $ac > b^2$.
If we had $|b| > a$ and $|b| > c$, then $b^2 > ac$, a contradiction. So $\max\{a, c\} \geqslant |b|$: an entry of largest absolute value lies on the diagonal.

9.6 Bilinear and Quadratic Forms

Inner products are symmetric bilinear forms with an extra positivity condition. We now drop positivity and classify symmetric bilinear forms themselves. Throughout this section, V is a finite-dimensional vector space over a field k with $\operatorname{char} k \neq 2$ (characteristic 2 appears only in a note), and B is a symmetric bilinear form on V.

9.6.1 Quadratic Forms and Polarization

Definition 9.6.1 (Quadratic form)

A ***quadratic form*** on V is a function $Q : V \to k$ of the form $Q(v) = B(v, v)$ for some bilinear form B on V. If $\{e_1, \dots, e_n\}$ is a basis and $A = (a_{ij})$ is the matrix of B, then

$$Q\Big(\sum_i x_i e_i\Big) = \sum_{i,j} a_{ij} x_i x_j = x^{t} A x,$$

a homogeneous polynomial of degree 2 in the coordinates.

Proposition 9.6.2 (Polarization)

Let $\operatorname{char} k \neq 2$. If B is a symmetric bilinear form and $Q(v) = B(v, v)$, then

$$B(u, v) = \tfrac{1}{2}\big(Q(u + v) - Q(u) - Q(v)\big) \quad \text{for all } u, v \in V.$$

Consequently, $B \mapsto Q$ is a bijection from symmetric bilinear forms on V to quadratic forms on V.

Proof. Expanding, $Q(u + v) = B(u, u) + B(u, v) + B(v, u) + B(v, v) = Q(u) + Q(v) + 2B(u, v)$, and we may divide by 2. So Q determines B, and the map is injective. For surjectivity, let $Q(v) = B'(v, v)$ for some bilinear form B'. Then $B(u, v) = \frac{1}{2}(B'(u, v) + B'(v, u))$ is a symmetric bilinear form with $B(v, v) = B'(v, v) = Q(v)$. □

Example 9.6.3 (A quadratic form and its matrix)

On k^2, the quadratic form $Q(x, y) = x^2 + 4xy - y^2$ corresponds to the symmetric matrix $\left(\begin{smallmatrix} 1 & 2 \\ 2 & -1 \end{smallmatrix}\right)$, i.e., to the symmetric bilinear form

$$B((x, y), (x', y')) = xx' + 2xy' + 2yx' - yy'.$$

The cross term $4xy$ is split equally between the two off-diagonal entries.

NOTE In characteristic 2, polarization fails, and so does the classification below. Over $\mathbb{F}_2$, the form $B(x, y) = x_1 y_2 + x_2 y_1$ on $\mathbb{F}_2^2$ is symmetric and nonzero, but $B(v, v) = 2x_1 x_2 = 0$ for every v, so it has the same quadratic form as the zero form. It also has no orthogonal basis in the sense of

Theorem 9.6.6 below: in such a basis its matrix would be diagonal with entries $B(e_i, e_i) = 0$, i.e., zero.

9.6.2 Orthogonal Bases

To classify symmetric bilinear forms, we look for a vector v with $B(v, v) \neq 0$ and then pass to $\operatorname{span}(v)^{\perp}$, as in the inductive construction of orthonormal bases.

Lemma 9.6.4 (Anisotropic vectors exist)

If $\operatorname{char} k \neq 2$ and $B \neq 0$ is symmetric, then there is $v \in V$ with $B(v, v) \neq 0$.

Proof. Choose u, w with $B(u, w) \neq 0$. By polarization, $2B(u, w) = Q(u + w) - Q(u) - Q(w) \neq 0$, so at least one of $Q(u + w)$, $Q(u)$, $Q(w)$ is nonzero. □

Lemma 9.6.5 (Splitting off a vector)

If $B(v, v) \neq 0$, then $V = \operatorname{span}(v) \oplus v^{\perp}$, where $v^{\perp} = \{w \in V : B(w, v) = 0\}$.

Proof. The linear functional $B(v, \cdot)$ is nonzero (its value at v is $B(v, v) \neq 0$), and since B is symmetric its kernel is $v^{\perp}$, which therefore has dimension $\dim V - 1$. If $cv \in v^{\perp}$, then $cB(v, v) = 0$, so $c = 0$; hence $\operatorname{span}(v) \cap v^{\perp} = \{0\}$, and the dimensions add up to $\dim V$. Explicitly, $w = \frac{B(v,w)}{B(v,v)} v + \left(w - \frac{B(v,w)}{B(v,v)} v\right)$. □

Theorem 9.6.6 (Orthogonal bases for symmetric forms)

Let $\operatorname{char} k \neq 2$ and let B be a symmetric bilinear form on a finite-dimensional space V. Then V has an ***orthogonal basis*** for B: a basis $\{e_1, \dots, e_n\}$ with $B(e_i, e_j) = 0$ for $i \neq j$. Equivalently, for every symmetric matrix $A \in \mathcal{M}_{n \times n}(k)$ there is an invertible P such that $P^t A P$ is diagonal.

Proof. Induction on $n = \dim V$. If $n = 0$ or $B = 0$, any basis works. Otherwise, Lemma 9.6.4 gives e_1 with $B(e_1, e_1) \neq 0$, and Lemma 9.6.5 gives $V = \operatorname{span}(e_1) \oplus W$ with $W = e_1^{\perp}$ of dimension $n - 1$. The restriction of B to W is symmetric and bilinear, so by induction W has a basis $\{e_2, \dots, e_n\}$ with $B(e_i, e_j) = 0$ for $2 \leqslant i \neq j$. Also $B(e_j, e_1) = B(e_1, e_j) = 0$ for $j \geqslant 2$, since $e_j \in W$. So $\{e_1, \dots, e_n\}$ is an orthogonal basis.

For the matrix statement, apply this to the form $B(x, y) = x^t A y$ on k^n and let P be the matrix whose columns are the e_i; by Proposition 9.1.6, $P^t A P$ is the (diagonal) matrix of B in the basis $\{e_i\}$. □

NOTE In an orthogonal basis the matrix of B is $\operatorname{diag}(d_1, \dots, d_n)$ with $d_i = B(e_i, e_i)$. So the rank of B is the number of nonzero d_i, and B is nondegenerate iff every $d_i \neq 0$. The d_i themselves are not determined by B: replacing e_i by ce_i ($c \neq 0$) replaces d_i by $c^2 d_i$. The classification problem is to decide how far the d_i can be normalized, and this depends on which elements of k are squares.

9.6.3 Classification over $\mathbb{C}$ and over $\mathbb{R}$

To state the classification, we say that a bilinear form B on V and a bilinear form B' on V' are ***equivalent*** if there is a linear isomorphism $T : V \to V'$ with $B'(T(u), T(v)) = B(u, v)$ for all $u, v \in V$. By Proposition 9.1.6, two forms on the same space are equivalent iff their matrices in a fixed basis are congruent; and every form on an n-dimensional space is equivalent, via a coordinate isomorphism, to the form $x^t A y$ on k^n, where A is its matrix in some basis.

Over $\mathbb{C}$, every number is a square, so everything can be normalized.

Theorem 9.6.7 (Symmetric forms over $\mathbb{C}$)

Let B be a nondegenerate symmetric bilinear form on a finite-dimensional complex vector space V. Then V has a basis $\{e_1, \dots, e_n\}$ with

$$B(e_i, e_j) = \delta_{ij},$$

i.e., the matrix of B is the identity. In particular, any two nondegenerate symmetric bilinear forms on complex vector spaces of the same finite dimension are equivalent.

Proof. Let $\{f_1, \dots, f_n\}$ be an orthogonal basis (Theorem 9.6.6) and $d_i = B(f_i, f_i)$, which is nonzero because B is nondegenerate. Choose $c_i \in \mathbb{C}$ with $c_i^2 = d_i$, and set $e_i = f_i / c_i$. Then $B(e_i, e_i) = d_i / c_i^2 = 1$, and $B(e_i, e_j) = 0$ for $i \neq j$. □

NOTE Over $\mathbb{C}$, symmetric bilinear forms have little to do with geometry: the form $B(z, w) = \sum z_i w_i$ on $\mathbb{C}^n$ (whose matrix is I) has $B(v, v) = 0$ for $v = (1, i, 0, \dots, 0)$. The forms that measure lengths over $\mathbb{C}$ are the Hermitian ones. The same proof shows that a possibly degenerate symmetric form on a complex space has, in a suitable basis, the matrix $\operatorname{diag}(I_r, 0)$, where r is its rank.

Over $\mathbb{R}$, only nonnegative numbers are squares, so the best we can do is normalize each d_i to ± 1.

Theorem 9.6.8 (Symmetric forms over $\mathbb{R}$)

Let B be a nondegenerate symmetric bilinear form on a real vector space V of dimension n. Then there is an integer p with $0 \leqslant p \leqslant n$ and a basis $\{e_1, \dots, e_n\}$ with $B(e_i, e_j) = 0$ for $i \neq j$, $B(e_i, e_i) = 1$ for $i \leqslant p$, and $B(e_i, e_i) = -1$ for $i > p$. That is,

$$B\Big(\sum_{i=1}^{n} x_i e_i, \sum_{i=1}^{n} y_i e_i\Big) = \sum_{i=1}^{p} x_i y_i - \sum_{i=p+1}^{n} x_i y_i, \qquad \text{matrix } \begin{pmatrix} I_p & 0 \\ 0 & -I_{n-p} \end{pmatrix}.$$

Proof. Let $\{f_1, \dots, f_n\}$ be an orthogonal basis with $d_i = B(f_i, f_i) \neq 0$, and set $e_i = f_i/\sqrt{|d_i|}$. Then $B(e_i, e_i) = d_i/|d_i| = \pm 1$. Reorder the basis so that the $+1$'s come first. □

The number p does not depend on the basis. This is *Sylvester's law of inertia*.

Theorem 9.6.9 (Sylvester's law of inertia)

In Theorem 9.6.8, p is the largest dimension of a subspace $W \subseteq V$ on which B is positive-definite ($B(w, w) > 0$ for all nonzero $w \in W$), and $n - p$ is the largest dimension of a subspace on which B is negative-definite. In particular, p is the same for every basis as in Theorem 9.6.8.

Proof. Let

$$V_+ = \operatorname{span}(e_1, \dots, e_p), \qquad V_- = \operatorname{span}(e_{p+1}, \dots, e_n).$$

For $v = \sum_{i \leqslant p} x_i e_i \in V_+$, $B(v, v) = \sum_{i \leqslant p} x_i^2 > 0$ unless $v = 0$, so B is positive-definite on V_+, a subspace of dimension p.

Now let W be any subspace on which B is positive-definite. For $v = \sum_{i > p} x_i e_i \in V_-$, $B(v, v) = -\sum_{i > p} x_i^2 \leqslant 0$. A nonzero vector of $W \cap V_-$ would satisfy both $B(v, v) > 0$ and $B(v, v) \leqslant 0$; so $W \cap V_- = \{0\}$. Hence

$$\dim W + (n - p) = \dim W + \dim V_- = \dim(W + V_-) \leqslant n,$$

and $\dim W \leqslant p$. So p is the largest dimension of a positive-definite subspace, which is defined without reference to any basis. Applying this to the form $-B$, whose basis as in Theorem 9.6.8 is $e_{p+1}, \dots, e_n, e_1, \dots, e_p$, shows that $n - p$ is the largest dimension of a negative-definite subspace. □

Definition 9.6.10 (Signature)

The ***signature*** of a nondegenerate symmetric bilinear form B on a real vector space of dimension n is the pair (p, q) with $q = n - p$, where p and q are as in Sylvester's law of inertia.

Corollary 9.6.11 (Classification over $\mathbb{R}$)

Let B and B' be nondegenerate symmetric bilinear forms on real vector spaces V and V' of the same finite dimension. Then B and B' are equivalent if and only if they have the same signature.

Proof. If both have signature (p, q), take bases $\{e_i\}$ of V for B and $\{e'_i\}$ of V' for B' as in Theorem 9.6.8, and let $T : V \to V'$ be the linear isomorphism with $T(e_i) = e'_i$; then $B'(T(e_i), T(e_j)) = B(e_i, e_j)$ for all i, j, hence $B'(T(u), T(v)) = B(u, v)$ for all u, v by bilinearity. Conversely, if T is an isomorphism with $B'(T(u), T(v)) = B(u, v)$, it maps subspaces on which B is positive-definite to subspaces of the same dimension on which B' is positive-definite, and T^{-1} does the reverse; so the largest such dimensions agree. $\square$

Example 9.6.12 (Signatures)

- An inner product on $\mathbb{R}^n$ has signature $(n, 0)$; conversely, a form of signature $(n, 0)$ is positive-definite. So Theorem 9.6.8 with $p = n$ recovers the existence of orthonormal bases.
- The Minkowski form $v_1w_1 - v_2w_2 - v_3w_3 - v_4w_4$ on $\mathbb{R}^4$ has signature $(1, 3)$.
- The quadratic form $Q(x, y) = xy$ on $\mathbb{R}^2$ comes from $B((x, y), (x', y')) = \frac{1}{2}(xy' + yx')$. In the basis $e_1 = (1, 1)$, $e_2 = (1, -1)$, we have $B(e_1, e_1) = 1$, $B(e_2, e_2) = -1$, and $B(e_1, e_2) = \frac{1}{2}(-1 + 1) = 0$. So the signature is $(1, 1)$; in coordinates, $xy = \left(\frac{x+y}{2}\right)^2 - \left(\frac{x-y}{2}\right)^2$.

NOTE Two practical remarks. First, by the real spectral theorem, a real symmetric matrix A can be diagonalized by a matrix P with $P^t = P^{-1}$, so $P^tAP = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ is simultaneously the matrix of the operator $x \mapsto Ax$ and of the form x^tAy in the new basis. Hence the signature of x^tAy is (number of positive eigenvalues of A, number of negative eigenvalues of A). Second, since $\det(P^tAP) = \det(P)^2 \det(A)$, the sign of $\det A$ is $(-1)^q$. For degenerate forms the same argument as in Sylvester's law shows that

the numbers p and q of $+1$'s and -1's in a diagonal form are well defined, with $p + q = \operatorname{rank} B$.

9.6.4 Forms over the Rationals

Over $\mathbb{Q}$, much less can be normalized, because few rational numbers are squares, and number theory enters. The following example shows that the values a form takes can distinguish two forms that look alike.

Example 9.6.13 (Two forms over $\mathbb{Q}$)

Let B and B' be the forms on $\mathbb{Q}^2$ with matrices

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix},$$

so $Q(v) = v_1^2 + v_2^2$ and $Q'(v) = 2v_1^2 + v_2^2$. Both are nondegenerate, and over $\mathbb{R}$ both have signature $(2, 0)$.

Q' takes the value 3: $Q'(1, 1) = 3$.

Q does not take the value 3: suppose $v_1^2 + v_2^2 = 3$ with $v \in \mathbb{Q}^2$. Clearing denominators, $n_1^2 + n_2^2 = 3m^2$ with integers n_1, n_2, m, $m \neq 0$, and we may assume they have no common factor. Squares of integers are 0 or $1 \bmod 4$, so $n_1^2 + n_2^2 \equiv 0, 1, 2 \pmod 4$, while $3m^2 \equiv 0, 3 \pmod 4$. Hence both sides are $\equiv 0 \pmod 4$. Then m is even (an odd m would give $3m^2 \equiv 3$), and $n_1^2 + n_2^2 \equiv 0 \pmod 4$ forces n_1 and n_2 to be even. This contradicts the assumption that they have no common factor.

So B and B' are not equivalent over $\mathbb{Q}$: if $T : \mathbb{Q}^2 \to \mathbb{Q}^2$ were a linear isomorphism with $B(T(u), T(v)) = B'(u, v)$, then $Q(T(1, 1)) = Q'(1, 1) = 3$, which is impossible.

9.7 Symplectic Forms

We now turn to the opposite extreme from inner products: forms for which every vector is orthogonal to itself.

Definition 9.7.1 (Symplectic form)

A ***symplectic form*** on a vector space V over a field k is a nondegenerate alternating bilinear form B on V. When $\operatorname{char} k \neq 2$, alternating forms are the

same as skew-symmetric ones, so a symplectic form is just a nondegenerate skew-symmetric bilinear form.

Example 9.7.2 (The standard symplectic form)

The form $b((x_1,x_2),(y_1,y_2)) = x_1y_2 - x_2y_1$ on k^2 (Example 9.1.4) is symplectic. More generally, on k^{2n} with coordinates $(x_1,y_1,\dots,x_n,y_n)$, the ***standard symplectic form*** is

$$B\big((x,y),(x',y')\big) = \sum_{i=1}^{n} (x_iy_i' - y_ix_i').$$

Its matrix is block diagonal,

$$\Omega_{2n} = \begin{pmatrix} \Omega_2 & & \\ & \ddots & \\ & & \Omega_2 \end{pmatrix}, \qquad \Omega_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

It is alternating (the terms of $B(v,v)$ cancel in pairs) and nondegenerate ($\det \Omega_{2n} = 1$).

NOTE When $\operatorname{char} k \neq 2$, there is a quick reason why symplectic forms exist only in even dimension: the matrix A of a skew-symmetric form on a space of dimension m satisfies $\det A = \det A^t = \det(-A) = (-1)^m \det A$, so if m is odd then $2\det A = 0$ and $\det A = 0$. The next theorem proves much more, in every characteristic.

Theorem 9.7.3 (Symplectic bases)

Let B be a symplectic form on a finite-dimensional vector space V over any field k. Then $\dim V = 2n$ is even, and V has a basis $(e_1,f_1,\dots,e_n,f_n)$ such that

$$B(e_i,e_j) = B(f_i,f_j) = 0, \qquad B(e_i,f_j) = \delta_{ij} = -B(f_j,e_i)$$

for all i,j. In such a ***symplectic basis***, the matrix of B is Ω_{2n}.

Proof. Induction on $\dim V$. If $V = 0$, take $n = 0$ and the empty basis. Otherwise pick any nonzero $e_1 \in V$. Since B is nondegenerate, the functional $B(e_1,\cdot)$ is nonzero, so there is $f \in V$ with $B(e_1,f) \neq 0$; set $f_1 = f/B(e_1,f)$, so that $B(e_1,f_1) = 1$,

and $B(f_1, e_1) = -1$ because alternating forms are skew-symmetric. The vectors e_1, f_1 are linearly independent: if $f_1 = ce_1$, then $B(e_1, f_1) = cB(e_1, e_1) = 0$.

Let $U = \operatorname{span}(e_1, f_1)$. Then $U \cap U^\perp = \{0\}$: if $v = ae_1 + bf_1$ satisfies $B(v, e_1) = B(v, f_1) = 0$, then $0 = B(v, e_1) = -b$ and $0 = B(v, f_1) = a$. Since B is nondegenerate, $\dim U^\perp = \dim V - 2$ (Lemma 9.1.30), so

$$V = U \oplus U^\perp.$$

The restriction of B to $U^\perp$ is alternating, and it is nondegenerate: if $w \in U^\perp$ satisfies $B(w, w') = 0$ for all $w' \in U^\perp$, then, since also $B(w, u) = 0$ for all $u \in U$, $B(w, \cdot)$ vanishes on $U + U^\perp = V$, so $w = 0$. By induction, $\dim U^\perp = 2(n-1)$ is even, and $U^\perp$ has a basis $(e_2, f_2, \dots, e_n, f_n)$ with the required relations among themselves. The remaining relations $B(e_1, x) = B(f_1, x) = 0$ for $x \in \{e_j, f_j\}$, $j \geqslant 2$, hold because $x \in U^\perp$ and B is skew-symmetric. So $(e_1, f_1, \dots, e_n, f_n)$ is a symplectic basis, and $\dim V = 2n$. □

Corollary 9.7.4 (Uniqueness of symplectic forms)

Any two symplectic forms on vector spaces of the same finite dimension are equivalent. In matrix terms, if $\operatorname{char} k \neq 2$, every invertible skew-symmetric matrix $A \in \mathcal{M}_{2n\times 2n}(k)$ satisfies $P^t A P = \Omega_{2n}$ for some invertible P.

Proof. Map a symplectic basis of one space to a symplectic basis of the other. For the matrix statement, apply the theorem to $B(x, y) = x^t A y$ on k^{2n} and use Proposition 9.1.6, with P the matrix whose columns are the symplectic basis. □

NOTE Compare with symmetric forms: over $\mathbb{R}$ they are classified by the signature, and over $\mathbb{Q}$ the classification involves number theory. Nondegenerate alternating forms, by contrast, have a single normal form over every field, determined by the dimension alone.

Definition 9.7.5 (Symplectic group)

Let B be a symplectic form on V with $\dim V = 2n$. The ***symplectic group*** of (V, B) is

$$\operatorname{Sp}(V, B) = \{T \in \operatorname{GL}(V) : B(T(u), T(v)) = B(u, v) \text{ for all } u, v \in V\}.$$

The ***symplectic group*** $\operatorname{Sp}(2n, k)$ is $\{A \in \operatorname{GL}_{2n}(k) : A^t \Omega_{2n} A = \Omega_{2n}\}$.

Proposition 9.7.6 (The symplectic group)

$\mathrm{Sp}(V, B)$ is a subgroup of $\mathrm{GL}(V)$, and any linear $T : V \to V$ with $B(T(u), T(v)) = B(u, v)$ for all u, v is automatically invertible. Choosing a symplectic basis identifies $\mathrm{Sp}(V, B)$ with $\mathrm{Sp}(2n, k)$.

Proof. If $T(u) = 0$, then $B(u, v) = B(T(u), T(v)) = 0$ for all v, so $u = 0$; thus T is injective, hence invertible. The identity preserves B; composites of maps preserving B preserve B; and $B(T^{-1}(u), T^{-1}(v)) = B(T(T^{-1}(u)), T(T^{-1}(v))) = B(u, v)$. Finally, if M is the matrix of T in a symplectic basis and x, y are coordinate vectors, then $B(T(u), T(v)) = (Mx)^t \Omega_{2n} (My) = x^t M^t \Omega_{2n} M y$, and this equals $x^t \Omega_{2n} y$ for all x, y iff $M^t \Omega_{2n} M = \Omega_{2n}$. So $T \mapsto M$, which is an isomorphism $\mathrm{GL}(V) \cong \mathrm{GL}_{2n}(k)$, carries $\mathrm{Sp}(V, B)$ onto $\mathrm{Sp}(2n, k)$. □

Example 9.7.7 ($\mathrm{Sp}(2, k) = \mathrm{SL}_2(k)$)

For $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, a direct computation gives

$$A^t \Omega_2 A = \begin{pmatrix} a & c \\ b & d \end{pmatrix} \begin{pmatrix} c & d \\ -a & -b \end{pmatrix} = \begin{pmatrix} 0 & ad - bc \\ -(ad - bc) & 0 \end{pmatrix} = \det(A)\, \Omega_2.$$

So $A \in \mathrm{Sp}(2, k)$ iff $\det A = 1$. Geometrically, over $\mathbb{R}$, a linear map of the plane preserves the signed-area form iff it preserves signed area. In every dimension, taking determinants in $A^t \Omega_{2n} A = \Omega_{2n}$ gives $\det(A)^2 = 1$ for $A \in \mathrm{Sp}(2n, k)$.

9.7.1 Hermitian Forms as Real Forms

A complex vector space can be regarded as a real vector space of twice the dimension. Under this change of viewpoint, a Hermitian form splits into a symmetric form and an alternating form, and the unitary group appears as an intersection of real groups.

Let V be a finite-dimensional complex vector space, and let $V_{\mathbb{R}}$ denote V regarded as a real vector space. If $\{e_1, \dots, e_n\}$ is a basis of V over $\mathbb{C}$, then $\{e_1, ie_1, \dots, e_n, ie_n\}$ is a basis of $V_{\mathbb{R}}$ over $\mathbb{R}$ (since $\sum (a_j + ib_j) e_j = \sum (a_j e_j + b_j\, ie_j)$ for real a_j, b_j), so $\dim_{\mathbb{R}} V_{\mathbb{R}} = 2n$. Let $J : V_{\mathbb{R}} \to V_{\mathbb{R}}$ be multiplication by i; it is $\mathbb{R}$-linear with $J^2 = -\mathrm{id}$. An $\mathbb{R}$-linear map $T : V_{\mathbb{R}} \to V_{\mathbb{R}}$ is $\mathbb{C}$-linear iff $T \circ J = J \circ T$, since $(a + ib)v = av + bJ(v)$.

Proposition 9.7.8 (Decomposition of Hermitian forms)

Let H be a Hermitian form on V. Define $G, B : V_{\mathbb{R}} \times V_{\mathbb{R}} \to \mathbb{R}$ by

$$G(u,v) = \operatorname{Re} H(u,v), \qquad B(u,v) = \operatorname{Im} H(u,v),$$

so that $H = G + iB$. Then:

(i) G is a symmetric bilinear form on $V_{\mathbb{R}}$.

(ii) B is an alternating bilinear form on $V_{\mathbb{R}}$.

(iii) $G(Ju, Jv) = G(u,v)$, $B(Ju, Jv) = B(u,v)$, $B(u,v) = G(Ju,v)$, and $G(u,v) = B(u, Jv)$ for all u, v.

(iv) H is nondegenerate iff G is nondegenerate iff B is nondegenerate.

(v) If H is a Hermitian inner product, then G is an inner product on $V_{\mathbb{R}}$ and B is a symplectic form on $V_{\mathbb{R}}$.

Proof. (i) H is additive in each variable, and $H(au, v) = aH(u,v) = H(u, av)$ for real a; since Re is $\mathbb{R}$-linear, G is $\mathbb{R}$-bilinear. Symmetry: $G(v,u) = \operatorname{Re}\overline{H(u,v)} = \operatorname{Re} H(u,v) = G(u,v)$.

(ii) Similarly B is $\mathbb{R}$-bilinear, and $B(v,v) = \operatorname{Im} H(v,v) = 0$ since $H(v,v)$ is real.

(iii) $H(iu, iv) = \bar{i}\, i\, H(u,v) = H(u,v)$, which gives the first two identities. Next, $H(iu, v) = -iH(u,v)$, so $G(Ju, v) = \operatorname{Re}(-iH(u,v)) = \operatorname{Im} H(u,v) = B(u,v)$. Finally $H(u, iv) = iH(u,v)$, so $B(u, Jv) = \operatorname{Im}(iH(u,v)) = \operatorname{Re} H(u,v) = G(u,v)$.

(iv) Replacing v by Jv in $G(u,v) = B(u, Jv)$ gives $G(u, Jv) = B(u, J^2 v) = -B(u,v)$. Suppose H is nondegenerate. If $G(u, \cdot) = 0$, then $B(u,v) = -G(u, Jv) = 0$ for all v, so $H(u, \cdot) = 0$ and $u = 0$. If $B(u, \cdot) = 0$, then $G(u,v) = B(u, Jv) = 0$ for all v, so again $H(u, \cdot) = 0$ and $u = 0$. Conversely, if G is nondegenerate and $H(u, \cdot) = 0$, then $G(u, \cdot) = \operatorname{Re} H(u, \cdot) = 0$, so $u = 0$; the same argument works for B.

(v) $G(v,v) = H(v,v) > 0$ for $v \neq 0$, so G is an inner product, and B is alternating by (ii) and nondegenerate by (iv). □

Proposition 9.7.9 (Two out of three)

Let H be a nondegenerate Hermitian form on V, with G and B as above, and let $T : V_{\mathbb{R}} \to V_{\mathbb{R}}$ be $\mathbb{R}$-linear. Consider the conditions:

(a) T preserves G: $G(Tu, Tv) = G(u,v)$ for all u, v;

(b) T preserves B;

(c) T is $\mathbb{C}$-linear: $TJ = JT$.

Any two of them imply the third. Moreover, T preserves H iff (a) and (b) hold.

Proof. (a) and (c) imply (b): $B(T(u),T(v)) = G(JT(u),T(v)) = G(TJ(u),T(v)) = G(Ju,v) = B(u,v)$.

(b) and (c) imply (a): $G(T(u),T(v)) = B(T(u),JT(v)) = B(T(u),TJ(v)) = B(u,Jv) = G(u,v)$.

(a) and (b) imply (c): first, T is injective (if $T(u) = 0$, then $G(u,v) = G(T(u),T(v)) = 0$ for all v, so $u = 0$ since G is nondegenerate), hence surjective. For all u, v,

$$G(JT(u),T(v)) = B(T(u),T(v)) = B(u,v) = G(Ju,v) = G(TJ(u),T(v)).$$

So $G(JT(u) - TJ(u), x) = 0$ for every $x = T(v) \in V_{\mathbb{R}}$, and $JT(u) = TJ(u)$ by nondegeneracy of G.

The last statement holds because $H = G + iB$ with G and B real-valued. □

Corollary 9.7.10 (The unitary group as an intersection)

Identify $\mathbb{C}^n$ with $\mathbb{R}^{2n}$ by $(z_1,\dots,z_n) \mapsto (x_1,y_1,\dots,x_n,y_n)$, where $z_j = x_j + iy_j$. For the standard Hermitian inner product $H(z,w) = \sum \overline{z_j} w_j$, the form G is the standard inner product of $\mathbb{R}^{2n}$ and B is the standard symplectic form of Example 9.7.2. Regarding $GL_n(\mathbb{C})$ as the subgroup of $GL_{2n}(\mathbb{R})$ of matrices commuting with J,

$$\begin{aligned} U(n) &= O(2n) \cap Sp(2n,\mathbb{R}) \cap GL_n(\mathbb{C}) \\ &= O(2n) \cap GL_n(\mathbb{C}) = Sp(2n,\mathbb{R}) \cap GL_n(\mathbb{C}) = O(2n) \cap Sp(2n,\mathbb{R}). \end{aligned}$$

Proof. With $z_j = x_j + iy_j$ and $w_j = x'_j + iy'_j$,

$$\overline{z_j}\, w_j = (x_j - iy_j)(x'_j + iy'_j) = (x_j x'_j + y_j y'_j) + i(x_j y'_j - y_j x'_j),$$

so G is the dot product on $\mathbb{R}^{2n}$ and B is the standard symplectic form. Thus $O(2n)$ and $Sp(2n,\mathbb{R})$ are the groups of real-linear maps preserving G and B, respectively. The group $U(n)$ consists of the $\mathbb{C}$-linear maps preserving H, i.e., those satisfying (a), (b), and (c) of Proposition 9.7.9, and by that proposition any two of the three conditions suffice. □

Exercise 9.7.11 (Bilinear and symplectic forms)

(1) Find the signature of the quadratic form $Q(x,y,z) = xy + yz + zx$ on $\mathbb{R}^3$, by constructing an orthogonal basis as in Theorem 9.6.6.

(2) Let B be a symplectic form on a vector space V of dimension $2n$. A subspace $L \subseteq V$ is ***isotropic*** if $B(u,v) = 0$ for all $u, v \in L$, i.e., $L \subseteq L^{\perp}$. Show that every isotropic subspace has dimension at most n, and that isotropic subspaces of dimension n exist.

Solution. **(1)** The symmetric bilinear form is

$$B(u,w) = \tfrac{1}{2}\big(u_1w_2 + u_2w_1 + u_2w_3 + u_3w_2 + u_3w_1 + u_1w_3\big).$$

Take $e_1 = (1,1,0)$, with $Q(e_1) = 1 \neq 0$. Then $B(e_1,w) = \frac{1}{2}(w_1 + w_2 + 2w_3)$, so $e_1^{\perp} = \{w : w_1 + w_2 + 2w_3 = 0\}$. In this plane take $e_2 = (1,-1,0)$, with $Q(e_2) = -1$, and $e_3 = (1,1,-1)$, with $Q(e_3) = 1 - 1 - 1 = -1$. They are orthogonal:

$$B(e_2,e_3) = \tfrac{1}{2}\big(1 - 1 + 1 + 0 + 0 - 1\big) = 0.$$

So in the orthogonal basis $\{e_1,e_2,e_3\}$ the matrix of B is $\operatorname{diag}(1,-1,-1)$, and the signature is $(1,2)$. (Equivalently, the matrix $\frac{1}{2}\left(\begin{smallmatrix}0&1&1\\1&0&1\\1&1&0\end{smallmatrix}\right)$ has eigenvalues $1, -\frac{1}{2}, -\frac{1}{2}$.)

(2) By Lemma 9.1.30, $\dim L^{\perp} = 2n - \dim L$. If $L \subseteq L^{\perp}$, then $\dim L \leqslant 2n - \dim L$, so $\dim L \leqslant n$. In a symplectic basis $(e_1,f_1,\dots,e_n,f_n)$, the subspace $\operatorname{span}(e_1,\dots,e_n)$ is isotropic, since $B(e_i,e_j) = 0$ for all i,j, and has dimension n.

9.8 Tensor Products

Tensors are an odd concept, mostly because their definitions look so elusive. In our view, the easiest way to understand them is to treat them as atomic objects, taken *tantum et tale*, and to concentrate on the rules by which they combine. The basic construction takes two vector spaces V and W over a field k and produces a new vector space $V \otimes W$, their ***tensor product***, built so that bilinear maps out of $V \times W$ become linear maps out of $V \otimes W$. We give three definitions, from the most concrete to the most abstract, and then prove that they agree.

Definition 9.8.1 (Bilinear map)

Let V, W and U be vector spaces over k. A map $\alpha : V \times W \to U$ is ***bilinear*** if it is linear in each variable when the other is held fixed:

$$\alpha(\lambda v + v', w) = \lambda\,\alpha(v,w) + \alpha(v',w), \qquad \alpha(v, \lambda w + w') = \lambda\,\alpha(v,w) + \alpha(v,w')$$

for all $\lambda \in k$, $v, v' \in V$ and $w, w' \in W$. A bilinear form on V (Definition 9.1.1) is the case $W = V$, $U = k$.

9.8.1 Three Definitions

Definition 9.8.2 (The tensor product via bases)

Let V and W be finite-dimensional, with bases $e_1, \dots, e_m$ of V and $f_1, \dots, f_n$ of W. Then $V \otimes W$ is the vector space with basis the mn symbols $e_i \otimes f_j$ ($1 \leqslant i \leqslant m$, $1 \leqslant j \leqslant n$); in particular $\dim(V \otimes W) = mn$. For $v = \sum_i a_i e_i \in V$ and $w = \sum_j b_j f_j \in W$ we set

$$v \otimes w = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j \, (e_i \otimes f_j).$$

The map $\beta : V \times W \to V \otimes W$, $\beta(v, w) = v \otimes w$, is bilinear.

Definition 9.8.3 (Pure tensors and rank)

Elements of $V \otimes W$ of the form $v \otimes w$ are called ***pure tensors***. Every element of $V \otimes W$ is a sum of pure tensors, since it is a linear combination of the pure tensors $e_i \otimes f_j$, and $c\,(e_i \otimes f_j) = (ce_i) \otimes f_j$ is again pure. The ***rank*** of $\eta \in V \otimes W$ is the smallest number r such that η is a sum of r pure tensors. Thus 0 has rank 0, and the nonzero pure tensors are exactly the tensors of rank 1.

Most tensors are not pure. For instance, if $m, n \geqslant 2$, then $e_1 \otimes f_1 + e_2 \otimes f_2$ is not pure: if it were equal to $(\sum_i a_i e_i) \otimes (\sum_j b_j f_j)$, comparing coefficients would give $a_1 b_1 = a_2 b_2 = 1$ and $a_1 b_2 = 0$, which is impossible. We return to tensor rank in Proposition 9.8.11.

Definition 9.8.2 is concrete, and it makes the dimension of $V \otimes W$ evident. It is also uncomfortable (Harris and Auroux, in their notes on tensor products, call it ugly): it depends on a choice of bases, and it is not clear that other bases give the same space. Our second definition chooses no bases at all.

Let F be the vector space with one basis vector $[v, w]$ for each pair $(v, w) \in V \times W$: its elements are the finite linear combinations of these symbols, with no relations among them. This space is enormous (when $k = \mathbb{R}$ and $V, W \neq 0$, its basis is uncountable), and in it the symbols $[v, w]$ are not bilinear in any sense. We force bilinearity by dividing out exactly the relations it requires. Let $F_0 \subseteq F$ be the subspace spanned by all elements

$$\begin{aligned} [\lambda v, w] - \lambda[v, w], \qquad & [v, \lambda w] - \lambda[v, w], \\ [v + v', w] - [v, w] - [v', w], \qquad & [v, w + w'] - [v, w] - [v, w'], \end{aligned}$$

for $\lambda \in k$, $v, v' \in V$ and $w, w' \in W$.

Definition 9.8.4 (The tensor product as a quotient)

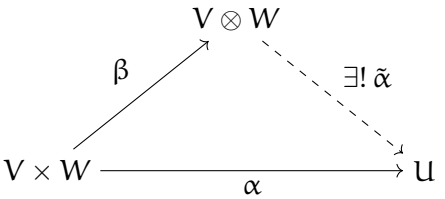
$V \otimes W$ is the quotient space F/F_0, and $v \otimes w$ denotes the class of $[v, w]$. The map $\beta : V \times W \to V \otimes W$, $\beta(v, w) = v \otimes w$, is bilinear, because that is precisely what the relations in F_0 say.

NOTE Nothing in this construction uses bases or finite dimension. With the same relations it also applies to modules over a commutative ring R, which are defined in Chapter 10 (Definition 10.1.1), where the result is written $M \otimes_R N$.

The most abstract characterization says what $V \otimes W$ does, without saying how it is built.

Definition 9.8.5 (The tensor product by its universal property)

A ***tensor product*** of V and W is a vector space $V \otimes W$ together with a bilinear map $\beta : V \times W \to V \otimes W$ with the following property: for every vector space U and every bilinear map $\alpha : V \times W \to U$, there is a *unique* linear map $\tilde{\alpha} : V \otimes W \to U$ with $\alpha = \tilde{\alpha} \circ \beta$. We write $v \otimes w = \beta(v, w)$.



9.8.2 The Universal Property

Theorem 9.8.6 (Existence and uniqueness of the tensor product)

Let V and W be finite-dimensional vector spaces over k.

(i) The space of Definition 9.8.2, with $\beta(v, w) = v \otimes w$, satisfies the universal property of Definition 9.8.5.

(ii) The space F/F_0 of Definition 9.8.4, with $\beta(v, w)$ the class of $[v, w]$, satisfies it as well.

(iii) If (Z, β) and (Z', β') both satisfy the universal property, there is a unique linear isomorphism $\theta : Z \to Z'$ with $\theta \circ \beta = \beta'$.

Consequently the three definitions produce the same space up to a unique isomorphism matching the pure tensors $v \otimes w$; in particular, the spaces that Definition 9.8.2 builds from different bases are canonically isomorphic. For any bases (e_i) of V and (f_j) of W, the pure tensors $e_i \otimes f_j$ form a basis of $V \otimes W$, and $\dim(V \otimes W) = (\dim V)(\dim W)$.

Proof. (i) Let $\alpha : V \times W \to U$ be bilinear, and let $\tilde{\alpha}$ be the linear map with $\tilde{\alpha}(e_i \otimes f_j) = \alpha(e_i, f_j)$ (a linear map may take arbitrary values on a basis). For $v = \sum_i a_i e_i$ and $w = \sum_j b_j f_j$, bilinearity gives

$$\alpha(v, w) = \sum_{i,j} a_i b_j \, \alpha(e_i, f_j) = \tilde{\alpha}\Big(\sum_{i,j} a_i b_j \, e_i \otimes f_j \Big) = \tilde{\alpha}(v \otimes w).$$

Any linear map ψ with $\psi \circ \beta = \alpha$ satisfies $\psi(e_i \otimes f_j) = \alpha(e_i, f_j)$, so $\psi = \tilde{\alpha}$, because a linear map is determined by its values on a basis.

(ii) Given α, let $\psi : F \to U$ be the linear map with $\psi([v, w]) = \alpha(v, w)$. Each spanning element of F_0 maps to 0 because α is bilinear; for example, $\psi([\lambda v, w] - \lambda[v, w]) = \alpha(\lambda v, w) - \lambda\alpha(v, w) = 0$. So $F_0 \subseteq \ker \psi$, and by the universal property of the quotient space (Theorem 7.1.7), ψ factors through a linear map $\tilde{\alpha} : F/F_0 \to U$ with $\tilde{\alpha}(v \otimes w) = \alpha(v, w)$. It is unique because the classes $v \otimes w$ span F/F_0.

(iii) The universal property of (Z, β), applied to the bilinear map β', gives a linear map $\theta : Z \to Z'$ with $\theta \circ \beta = \beta'$, and it is the only such map. Symmetrically there is $\theta' : Z' \to Z$ with $\theta' \circ \beta' = \beta$. Then $\theta' \circ \theta : Z \to Z$ is linear and $(\theta' \circ \theta) \circ \beta = \beta$; the identity of Z has the same property, so the uniqueness in the universal property of (Z, β), applied to $\alpha = \beta$, gives $\theta' \circ \theta = \mathrm{id}_Z$. Likewise $\theta \circ \theta' = \mathrm{id}_{Z'}$.

For the last statements: by (i)–(iii), any model Z of the tensor product is isomorphic to the space built in Definition 9.8.2 from the bases (e_i) and (f_j), by an isomorphism matching pure tensors. That isomorphism carries the basis $(e_i \otimes f_j)$ of the latter to the elements $e_i \otimes f_j$ of Z, which therefore form a basis of Z. □

NOTE The argument in (iii) applies to every object defined by a universal property, as for products and coproducts (Section 8.3): the universal property determines the object up to unique isomorphism, and a separate construction shows that it exists. From now on $V \otimes W$ means any of these spaces, and we use whichever description is convenient. In practice, two facts get used constantly: to define a linear map out of $V \otimes W$, it suffices to give a bilinear

map out of $V \times W$; and two linear maps out of $V \otimes W$ that agree on all pure tensors are equal, because pure tensors span.

NOTE The assignment $\alpha \mapsto \tilde{\alpha}$ is a bijection from the bilinear maps $V \times W \to U$ to the linear maps $V \otimes W \to U$, with inverse $\psi \mapsto \psi \circ \beta$. It is linear, so it identifies the vector space of bilinear maps $V \times W \to U$ with $\operatorname{Hom}(V \otimes W, U)$. In this sense the tensor product converts bilinear problems into linear ones.

9.8.3 Functoriality and Basic Identities

Proposition 9.8.7 (Tensor product of linear maps)

Let $\varphi : V \to V'$ and $\psi : W \to W'$ be linear maps between finite-dimensional vector spaces. There is a unique linear map $\varphi \otimes \psi : V \otimes W \to V' \otimes W'$ with

$$(\varphi \otimes \psi)(v \otimes w) = \varphi(v) \otimes \psi(w).$$

Moreover:

(i) $(\varphi' \circ \varphi) \otimes (\psi' \circ \psi) = (\varphi' \otimes \psi') \circ (\varphi \otimes \psi)$ for linear maps $\varphi' : V' \to V''$, $\psi' : W' \to W''$, and $\mathrm{id}_V \otimes \mathrm{id}_W = \mathrm{id}_{V \otimes W}$;

(ii) if $V' = V$ and $W' = W$, then $\operatorname{Tr}(\varphi \otimes \psi) = \operatorname{Tr}(\varphi)\,\operatorname{Tr}(\psi)$.

Proof. The map $V \times W \to V' \otimes W'$, $(v, w) \mapsto \varphi(v) \otimes \psi(w)$, is bilinear, because φ and ψ are linear and $\otimes$ is bilinear. By the universal property it induces a unique linear map $\varphi \otimes \psi$ with the stated values.

(i) On pure tensors,

$$((\varphi' \otimes \psi') \circ (\varphi \otimes \psi))(v \otimes w) = \varphi'(\varphi(v)) \otimes \psi'(\psi(w)) = ((\varphi' \circ \varphi) \otimes (\psi' \circ \psi))(v \otimes w),$$

and two linear maps that agree on pure tensors are equal. The same argument gives $\mathrm{id}_V \otimes \mathrm{id}_W = \mathrm{id}_{V \otimes W}$.

(ii) Let $A = (a_{ki})$ and $B = (b_{lj})$ be the matrices of φ and ψ in bases (e_i) of V and (f_j) of W. Then

$$(\varphi \otimes \psi)(e_i \otimes f_j) = \varphi(e_i) \otimes \psi(f_j) = \sum_{k,l} a_{ki} b_{lj}\, e_k \otimes f_l,$$

so the diagonal entry of the matrix of $\varphi \otimes \psi$ at $e_i \otimes f_j$ is $a_{ii} b_{jj}$. Hence $\operatorname{Tr}(\varphi \otimes \psi) = \sum_{i,j} a_{ii} b_{jj} = \operatorname{Tr}(A)\,\operatorname{Tr}(B)$. □

In the language of Definition 8.2.1, part (i) says that $\otimes$ is a functor in two variables: it is a functor in each variable when the other is held fixed, and the

two variables do not interfere. In particular, if φ and ψ are isomorphisms, so is $\varphi \otimes \psi$, with inverse $\varphi^{-1} \otimes \psi^{-1}$.

Proposition 9.8.8 (Basic identities)

Let U, V, W be finite-dimensional vector spaces over k. There are isomorphisms

(i) $V \otimes W \cong W \otimes V$, with $v \otimes w \mapsto w \otimes v$;

(ii) $(U \oplus V) \otimes W \cong (U \otimes W) \oplus (V \otimes W)$, with $(u, v) \otimes w \mapsto (u \otimes w, v \otimes w)$;

(iii) $k \otimes V \cong V$, with $\lambda \otimes v \mapsto \lambda v$.

Proof. In each case the formula comes from a bilinear map on the product of the two factors, namely $(v, w) \mapsto w \otimes v$, $((u, v), w) \mapsto (u \otimes w, v \otimes w)$ and $(\lambda, v) \mapsto \lambda v$. By the universal property each induces a linear map with the stated values on pure tensors. Each of these maps sends a basis to a basis, so it is an isomorphism. Indeed, let (g_l), (e_i), (f_j) be bases of U, V, W. In (i), $e_i \otimes f_j \mapsto f_j \otimes e_i$. In (ii), the vectors $(g_l, 0)$ and $(0, e_i)$ form a basis of $U \oplus V$, and $(g_l, 0) \otimes f_j \mapsto (g_l \otimes f_j, 0)$, $(0, e_i) \otimes f_j \mapsto (0, e_i \otimes f_j)$, which together form a basis of $(U \otimes W) \oplus (V \otimes W)$. In (iii), $1 \otimes e_i \mapsto e_i$. □

Example 9.8.9 (Complexification as a tensor product)

Let V be a finite-dimensional real vector space, and regard $\mathbb{C}$ as a real vector space with basis $1, i$. Write $V \otimes_{\mathbb{R}} \mathbb{C}$ for the tensor product of these two real vector spaces. For $\lambda \in \mathbb{C}$, multiplication by λ is an $\mathbb{R}$-linear map $m_\lambda : \mathbb{C} \to \mathbb{C}$, and we set $\lambda \cdot \xi = (\mathrm{id}_V \otimes m_\lambda)(\xi)$, so that $\lambda \cdot (v \otimes z) = v \otimes \lambda z$. Proposition 9.8.7(i) gives $(\lambda\mu) \cdot \xi = \lambda \cdot (\mu \cdot \xi)$ and $1 \cdot \xi = \xi$, and checking on pure tensors gives the distributive laws; so $V \otimes_{\mathbb{R}} \mathbb{C}$ is a complex vector space. If $e_1, \dots, e_n$ is a basis of V, the vectors $e_j \otimes 1$, $e_j \otimes i$ form a real basis of $V \otimes_{\mathbb{R}} \mathbb{C}$, and $e_j \otimes i = i \cdot (e_j \otimes 1)$, so the $e_j \otimes 1$ form a complex basis. The map from the complexification $V_{\mathbb{C}}$ of Definition 7.10.1,

$$V_{\mathbb{C}} \to V \otimes_{\mathbb{R}} \mathbb{C}, \qquad v + iw \mapsto v \otimes 1 + w \otimes i,$$

is $\mathbb{R}$-linear and sends a real basis to a real basis. It also commutes with multiplication by i: the vector $i(v + iw) = -w + iv$ goes to $-w \otimes 1 + v \otimes i = i \cdot (v \otimes 1 + w \otimes i)$. So it is an isomorphism of complex vector spaces, and under it $\varphi_{\mathbb{C}}$ corresponds to $\varphi \otimes \mathrm{id}_{\mathbb{C}}$.

9.8.4 Hom as a Tensor Product

The most useful identity involving tensor products is the following. For a basis $e_1, \dots, e_n$ of V we write $e_1^*, \dots, e_n^*$ for the dual basis of V^* (Definition 6.2.5), so $e_i^*(e_j) = \delta_{ij}$.

Proposition 9.8.10 ($\mathrm{Hom}(V,W) \cong V^* \otimes W$)
Let V and W be finite-dimensional. There is an isomorphism

$$\Phi : V^* \otimes W \xrightarrow{\sim} \mathrm{Hom}(V,W), \qquad \Phi(\ell \otimes w) = \big(v \mapsto \ell(v)\, w\big),$$

defined without any choice of bases. It is compatible with composition: for linear maps $\alpha : V' \to V$ and $\gamma : W \to W'$ between finite-dimensional spaces,

$$\gamma \circ \Phi(\eta) \circ \alpha = \Phi\big((\alpha^{\mathrm{t}} \otimes \gamma)(\eta)\big) \qquad \text{for all } \eta \in V^* \otimes W,$$

where $\alpha^{\mathrm{t}} : V^* \to V'^*$ is the dual map (Definition 6.2.9).

Proof. The map $V^* \times W \to \mathrm{Hom}(V,W)$, $(\ell, w) \mapsto (v \mapsto \ell(v)w)$, is bilinear: $(\lambda\ell + \ell')(v)\, w = \lambda\,\ell(v)w + \ell'(v)w$, and similarly in w. By the universal property it induces a linear map Φ with the stated values.

Choose bases $e_1, \dots, e_n$ of V and $f_1, \dots, f_m$ of W. The map $\Phi(e_i^* \otimes f_j)$ sends e_i to f_j and every other e_k to 0, so its matrix is the $m \times n$ matrix with a 1 in row j, column i, and zeros elsewhere. These mn matrices form a basis of the space of $m \times n$ matrices, which is isomorphic to $\mathrm{Hom}(V,W)$ by taking matrices in the chosen bases (Theorem 5.2.7). So Φ sends the basis $(e_i^* \otimes f_j)$ of $V^* \otimes W$ to a basis of $\mathrm{Hom}(V,W)$, and it is an isomorphism.

Both sides of the compatibility formula are linear in η, so it suffices to check it for $\eta = \ell \otimes w$. For $v' \in V'$,

$$\gamma\big(\Phi(\ell \otimes w)(\alpha(v'))\big) = \ell(\alpha(v'))\,\gamma(w) = (\alpha^{\mathrm{t}}\ell)(v')\,\gamma(w) = \Phi\big(\alpha^{\mathrm{t}}\ell \otimes \gamma(w)\big)(v'). \qquad \square$$

Under Φ, a pure tensor $\ell \otimes w$ with $\ell \neq 0$ and $w \neq 0$ is the map $v \mapsto \ell(v)w$, whose image is the line spanned by w: a map of rank one. This is the first case of a general fact, which explains the name "rank" for tensors.

Proposition 9.8.11 (Tensor rank equals rank)
Let V and W be finite-dimensional and $\eta \in V^* \otimes W$. The rank of η as a tensor equals the rank of the linear map $\Phi(\eta) : V \to W$.

Proof. Let r be the tensor rank of η and $\varphi = \Phi(\eta)$. If $\eta = \sum_{i=1}^{r} \ell_i \otimes w_i$, then $\varphi(v) = \sum_{i=1}^{r} \ell_i(v)\, w_i$ lies in $\mathrm{span}(w_1, \dots, w_r)$ for every v, so $\operatorname{rank}\varphi \leqslant r$.

Conversely, let $s = \operatorname{rank}\varphi$ and $n = \dim V$. By the Rank–Nullity Theorem (Theorem 6.1.12), $\dim\ker\varphi = n - s$. Choose a basis $e_{s+1}, \dots, e_n$ of $\ker\varphi$, extend it to a basis $e_1, \dots, e_n$ of V, and let (e_i^*) be the dual basis. The map $\Phi\big(\sum_{i=1}^{s} e_i^* \otimes \varphi(e_i)\big)$ sends e_j to $\varphi(e_j)$ for $j \leqslant s$ and to $0 = \varphi(e_j)$ for $j > s$, so it equals φ. Since Φ is injective, $\eta = \sum_{i=1}^{s} e_i^* \otimes \varphi(e_i)$, a sum of s pure tensors, and $r \leqslant s$. □

Example 9.8.12 (A 2×2 matrix as a tensor)

Let V and W have bases e_1, e_2 and f_1, f_2. The linear map with matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ sends $e_1 \mapsto af_1 + cf_2$ and $e_2 \mapsto bf_1 + df_2$, so it corresponds under Φ to the tensor

$$e_1^* \otimes (af_1 + cf_2) + e_2^* \otimes (bf_1 + df_2).$$

By Proposition 9.8.11, this tensor is pure exactly when the matrix has rank at most 1, that is, when its columns are proportional, that is, when $ad - bc = 0$. Otherwise it has rank 2.

Proposition 9.8.13 (Duals and bilinear forms)

Let V and W be finite-dimensional.

(i) There is an isomorphism $V^* \otimes W^* \to (V \otimes W)^*$ sending $\ell \otimes m$ to the linear functional with $v \otimes w \mapsto \ell(v)\, m(w)$.

(ii) The space $B(V)$ of bilinear forms on V is isomorphic to $V^* \otimes V^*$; the tensor $\ell \otimes m$ corresponds to the form $b(v, w) = \ell(v)\, m(w)$.

Proof. (i) For fixed (ℓ, m), the map $(v, w) \mapsto \ell(v)m(w)$ is bilinear, so it induces a linear functional on $V \otimes W$. Its values on pure tensors depend bilinearly on (ℓ, m), and pure tensors span, so the functional itself depends bilinearly on (ℓ, m); the universal property gives a linear map $V^* \otimes W^* \to (V \otimes W)^*$. With bases (e_i) of V and (f_j) of W, it sends $e_i^* \otimes f_j^*$ to the functional taking $e_k \otimes f_l$ to $\delta_{ik}\delta_{jl}$, which is the dual basis of the basis $(e_k \otimes f_l)$ of $V \otimes W$. So it sends a basis to a basis.

(ii) By the universal property, bilinear forms on V correspond to linear functionals on $V \otimes V$, and this correspondence is a linear isomorphism $B(V) \cong (V \otimes V)^*$. Combine it with (i) for $W = V$. □

NOTE Part (ii) agrees with the isomorphism $B(V) \cong \operatorname{Hom}(V, V^*)$, $b \mapsto \varphi_b$, of Theorem 9.1.8. Proposition 9.8.10 with $W = V^*$ gives $\operatorname{Hom}(V, V^*) \cong V^* \otimes V^*$, and the form $b(v, w) = \ell(v)m(w)$ has $\varphi_b(v) = b(v, \cdot) = \ell(v)\, m$, which

corresponds to $\ell \otimes m$. So both routes from $B(V)$ to $V^* \otimes V^*$ give the same isomorphism.

NOTE The dual map also has a description in terms of tensors. Chaining isomorphisms,

$$\mathrm{Hom}(V,W) \cong V^* \otimes W \cong W \otimes V^* \cong W^{**} \otimes V^* \cong \mathrm{Hom}(W^*, V^*),$$

using Proposition 9.8.10, the swap of Proposition 9.8.8(i), the isomorphism $W \cong W^{**}$ of Theorem 6.2.19 on the first factor, and Proposition 9.8.10 again. Follow a rank-one map $\varphi(v) = \ell(v)w$ through the chain: it goes to $\ell \otimes w$, then to $w \otimes \ell$, then to $\mathrm{ev}_w \otimes \ell$ (where $\mathrm{ev}_w(\alpha) = \alpha(w)$), and finally to the map $\alpha \mapsto \alpha(w)\,\ell$. That map is φ^t, because $\varphi^t(\alpha) = \alpha \circ \varphi = \alpha(w)\,\ell$. Both ends of the chain are linear in φ, and rank-one maps span $\mathrm{Hom}(V,W)$, so the composite sends every φ to its dual map φ^t. This is a construction of the transpose that uses no bases.

9.9 Multilinear Maps and the Tensor Algebra

9.9.1 Multilinear Maps

The constructions of the previous section extend from two factors to any finite number of them.

Definition 9.9.1 (Multilinear map)

Let $V_1, \dots, V_d$ and U be vector spaces over k. A map $m : V_1 \times \cdots \times V_d \to U$ is ***multilinear*** (or d-***linear***) if it is linear in each variable when the others are held fixed. When $U = k$, m is a ***multilinear form***.

Theorem 9.9.2 (Tensor product of several spaces)

Let $V_1, \dots, V_d$ be finite-dimensional, and let $\mathcal{B}_i$ be a basis of V_i. There is a vector space $V_1 \otimes \cdots \otimes V_d$ together with a multilinear map

$$\mu : V_1 \times \cdots \times V_d \to V_1 \otimes \cdots \otimes V_d, \qquad \mu(v_1, \dots, v_d) = v_1 \otimes \cdots \otimes v_d,$$

such that for every vector space U and every multilinear map $m : V_1 \times \cdots \times V_d \to U$ there is a unique linear map $\tilde{m} : V_1 \otimes \cdots \otimes V_d \to U$ with $m = \tilde{m} \circ \mu$. It is unique up to unique isomorphism, the pure tensors $b_1 \otimes \cdots \otimes b_d$ with $b_i \in \mathcal{B}_i$ form a basis of it, and $\dim(V_1 \otimes \cdots \otimes V_d) = \prod_i \dim V_i$.

Proof. This is the proof of Theorem 9.8.6 with more indices. Let $V_1 \otimes \cdots \otimes V_d$ be the vector space with basis the symbols $b_1 \otimes \cdots \otimes b_d$ ($b_i \in \mathcal{B}_i$). If $v_i = \sum_{b \in \mathcal{B}_i} a_{i,b}\, b$, set

$$\mu(v_1, \ldots, v_d) = \sum_{b_1, \ldots, b_d} a_{1,b_1} \cdots a_{d,b_d}\; b_1 \otimes \cdots \otimes b_d,$$

which is multilinear. Given m, the linear map with $\tilde{m}(b_1 \otimes \cdots \otimes b_d) = m(b_1, \ldots, b_d)$ satisfies $\tilde{m} \circ \mu = m$, by expanding $m(v_1, \ldots, v_d)$ using multilinearity; and it is the only linear map that can, since its values on a basis are forced. Uniqueness up to unique isomorphism is proved exactly as in Theorem 9.8.6(iii). □

As before, the space can also be built as a quotient of the vector space with basis all symbols $[v_1, \ldots, v_d]$, by the relations expressing multilinearity. Linear maps $\varphi_i : V_i \to W_i$ induce a linear map $\varphi_1 \otimes \cdots \otimes \varphi_d$ with $v_1 \otimes \cdots \otimes v_d \mapsto \varphi_1(v_1) \otimes \cdots \otimes \varphi_d(v_d)$, since the right side is multilinear in the v_i; exactly as in Proposition 9.8.7, this construction is compatible with composition and identities.

Proposition 9.9.3 (Associativity)

For finite-dimensional vector spaces $V_1, \ldots, V_a$ and $W_1, \ldots, W_b$, there is an isomorphism

$$V_1 \otimes \cdots \otimes V_a \otimes W_1 \otimes \cdots \otimes W_b \xrightarrow{\sim} (V_1 \otimes \cdots \otimes V_a) \otimes (W_1 \otimes \cdots \otimes W_b)$$

with $v_1 \otimes \cdots \otimes v_a \otimes w_1 \otimes \cdots \otimes w_b \mapsto (v_1 \otimes \cdots \otimes v_a) \otimes (w_1 \otimes \cdots \otimes w_b)$. In particular $(U \otimes V) \otimes W \cong U \otimes V \otimes W \cong U \otimes (V \otimes W)$.

Proof. The map $(v_1, \ldots, v_a, w_1, \ldots, w_b) \mapsto (v_1 \otimes \cdots \otimes v_a) \otimes (w_1 \otimes \cdots \otimes w_b)$ is multilinear: it is linear in v_i because $v_1 \otimes \cdots \otimes v_a$ is linear in v_i and $\otimes$ is linear in its first argument, and similarly for each w_j. By Theorem 9.9.2 it induces a linear map Ψ with the stated values. Choose bases of all the spaces. Then Ψ sends the basis tensor $b_1 \otimes \cdots \otimes b_a \otimes c_1 \otimes \cdots \otimes c_b$ to $(b_1 \otimes \cdots \otimes b_a) \otimes (c_1 \otimes \cdots \otimes c_b)$. The tensors $b_1 \otimes \cdots \otimes b_a$ form a basis of $V_1 \otimes \cdots \otimes V_a$, and similarly for the c's, so by Theorem 9.8.6 these images form a basis of the target. So Ψ sends a basis to a basis. □

Thus nothing new happens with more than two factors, and we may drop parentheses in iterated tensor products.

9.9.2 Tensor Powers and the Tensor Algebra

Definition 9.9.4 (Tensor powers)

For a vector space V and $d \geqslant 1$, the dth ***tensor power*** of V is

$$V^{\otimes d} = \underbrace{V \otimes \cdots \otimes V}_{d}.$$

By convention, $V^{\otimes 0} = k$ and $V^{\otimes 1} = V$. If $\dim V = n$, then $\dim V^{\otimes d} = n^d$. For a linear map $\varphi : V \to W$ we write $\varphi^{\otimes d} = \varphi \otimes \cdots \otimes \varphi : V^{\otimes d} \to W^{\otimes d}$.

By Proposition 9.9.3, $V^{\otimes a} \otimes V^{\otimes b} \cong V^{\otimes (a+b)}$. Composing this isomorphism with the bilinear map $V^{\otimes a} \times V^{\otimes b} \to V^{\otimes a} \otimes V^{\otimes b}$ gives a bilinear multiplication $V^{\otimes a} \times V^{\otimes b} \to V^{\otimes (a+b)}$,

$$(v_1 \otimes \cdots \otimes v_a) \cdot (w_1 \otimes \cdots \otimes w_b) = v_1 \otimes \cdots \otimes v_a \otimes w_1 \otimes \cdots \otimes w_b,$$

which is concatenation of tensors. When $a = 0$ or $b = 0$ it is multiplication by a scalar.

Definition 9.9.5 (Tensor algebra)

The ***tensor algebra*** of V is the direct sum

$$T(V) = \bigoplus_{d=0}^{\infty} V^{\otimes d} = k \oplus V \oplus (V \otimes V) \oplus (V \otimes V \otimes V) \oplus \cdots,$$

whose elements are finite sums $\eta_0 + \eta_1 + \cdots$ with $\eta_d \in V^{\otimes d}$, with the multiplication that extends the products $V^{\otimes a} \times V^{\otimes b} \to V^{\otimes (a+b)}$ bilinearly.

NOTE $T(V)$ is a ring: the multiplication distributes over addition because it is bilinear, it has the identity $1 \in V^{\otimes 0} = k$, and it is associative because $(xy)z$ and $x(yz)$ are trilinear in (x, y, z) and agree on pure tensors, where both are concatenation. It is also a vector space over k whose multiplication is bilinear; a ring with this extra structure is called a k-***algebra***. The tensor algebra is ***graded***: it is the direct sum of the subspaces $V^{\otimes d}$, and $V^{\otimes a} \cdot V^{\otimes b} \subseteq V^{\otimes (a+b)}$. It is not commutative once $\dim V \geqslant 2$: for linearly independent u, v, the tensors $u \otimes v$ and $v \otimes u$ are distinct elements of a basis of $V \otimes V$. As a vector space, $T(V)$ is infinite-dimensional unless $V = 0$.
The tensor algebra is the ***free*** k-***algebra*** on V: every linear map $f : V \to A$ into a k-algebra A extends uniquely to a k-linear ring homomorphism

$F : T(V) \to A$ with $F(1) = 1$, namely $F(v_1 \otimes \cdots \otimes v_d) = f(v_1) \cdots f(v_d)$. This formula defines F on each $V^{\otimes d}$, because the right side is multilinear in the v_i; F is multiplicative because $F(xy)$ and $F(x)F(y)$ are bilinear in (x, y) and agree on pure tensors; and F is unique because $T(V)$ is spanned by 1 and the products $v_1 \otimes \cdots \otimes v_d$ of elements of V.

Proposition 9.9.6 (Multilinear forms)

For finite-dimensional vector spaces $V_1, \ldots, V_d$ there are isomorphisms

$$V_1^* \otimes \cdots \otimes V_d^* \cong (V_1 \otimes \cdots \otimes V_d)^* \cong \{\text{multilinear forms on } V_1 \times \cdots \times V_d\},$$

under which $\ell_1 \otimes \cdots \otimes \ell_d$ corresponds to the form $(v_1, \ldots, v_d) \mapsto \ell_1(v_1) \cdots \ell_d(v_d)$. In particular $(V^*)^{\otimes d} \cong (V^{\otimes d})^*$ is the space of d-linear forms on V.

Proof. The second isomorphism is the universal property: $m \mapsto \tilde{m}$ is a linear bijection from multilinear forms to linear functionals on $V_1 \otimes \cdots \otimes V_d$. For the first, argue as in Proposition 9.8.13(i): the form $(v_1, \ldots, v_d) \mapsto \prod_i \ell_i(v_i)$ depends multilinearly on $(\ell_1, \ldots, \ell_d)$, so it induces a linear map $V_1^* \otimes \cdots \otimes V_d^* \to (V_1 \otimes \cdots \otimes V_d)^*$, and this map sends the tensor product of dual basis vectors $b_1^* \otimes \cdots \otimes b_d^*$ to the functional that is 1 on $b_1 \otimes \cdots \otimes b_d$ and 0 on the other basis tensors. So it sends a basis to the dual basis. □

9.10 Symmetric and Exterior Powers

The space $B(V) \cong V^* \otimes V^*$ of bilinear forms contains the symmetric and the alternating forms. This section carries out the analogous constructions for all tensor powers. Throughout, V is a finite-dimensional vector space over k, $n = \dim V$, and $d \geqslant 0$.

9.10.1 The Symmetric Group Acting on Tensors

For $\sigma \in S_d$, the map $V^d \to V^{\otimes d}$, $(v_1, \ldots, v_d) \mapsto v_{\sigma^{-1}(1)} \otimes \cdots \otimes v_{\sigma^{-1}(d)}$, is multilinear, so it induces a linear map $V^{\otimes d} \to V^{\otimes d}$, written $\eta \mapsto \sigma \cdot \eta$:

$$\sigma \cdot (v_1 \otimes v_2 \otimes \cdots \otimes v_d) = v_{\sigma^{-1}(1)} \otimes v_{\sigma^{-1}(2)} \otimes \cdots \otimes v_{\sigma^{-1}(d)}.$$

The factor in position i moves to position $\sigma(i)$. This is an action of S_d on $V^{\otimes d}$ by linear maps (Definition 3.2.1). Indeed, for $\sigma, \tau \in S_d$, writing $w_i = v_{\tau^{-1}(i)}$,

$$\sigma \cdot \big(\tau \cdot (v_1 \otimes \cdots \otimes v_d)\big) = \sigma \cdot (w_1 \otimes \cdots \otimes w_d) = w_{\sigma^{-1}(1)} \otimes \cdots \otimes w_{\sigma^{-1}(d)}$$

$$= v_{\tau^{-1}\sigma^{-1}(1)} \otimes \cdots \otimes v_{\tau^{-1}\sigma^{-1}(d)},$$

which is $(\sigma\tau)\cdot(v_1 \otimes \cdots \otimes v_d)$ because $(\sigma\tau)^{-1} = \tau^{-1}\sigma^{-1}$. Both sides are linear, so $\sigma \cdot (\tau \cdot \eta) = (\sigma\tau)\cdot\eta$ for all η, and clearly $e \cdot \eta = \eta$. (The inverses make this a left action for our convention that $\sigma\tau$ applies τ first.)

Definition 9.10.1 (Symmetric and alternating multilinear maps)

A multilinear map $m : V^d \to U$ (where $V^d = V \times \cdots \times V$) is ***symmetric*** if

$$m(v_{\sigma(1)}, \ldots, v_{\sigma(d)}) = m(v_1, \ldots, v_d) \qquad \text{for all } \sigma \in S_d \text{ and all } v_i \in V,$$

and ***alternating*** if $m(v_1, \ldots, v_d) = 0$ whenever $v_i = v_j$ for some $i \neq j$. For $U = k$, the alternating multilinear maps are the alternating forms of Definition 5.4.1, and for $d = 2$ and $U = k$ the symmetric and alternating ones are the symmetric and alternating bilinear forms of Definition 9.1.2.

Lemma 9.10.2 (Alternating maps change sign)

Let $m : V^d \to U$ be multilinear and alternating.

(i) Exchanging two of the arguments of m changes the sign of its value.

(ii) $m(v_{\sigma(1)}, \ldots, v_{\sigma(d)}) = \operatorname{sgn}(\sigma)\, m(v_1, \ldots, v_d)$ for every $\sigma \in S_d$.

(iii) $m(v_1, \ldots, v_d) = 0$ whenever $v_1, \ldots, v_d$ are linearly dependent.

Conversely, if $\operatorname{char} k \neq 2$, a multilinear map with property (i) is alternating.

Proof. For alternating multilinear forms, that is, for $U = k$, these are parts (a)–(c) of Lemma 5.4.2. Their proofs use only that the map is multilinear and alternating, never that its values are scalars, so they apply word for word to m. For instance, (i) comes from expanding

$$0 = m(\ldots, x+y, \ldots, x+y, \ldots) = m(\ldots, x, \ldots, y, \ldots) + m(\ldots, y, \ldots, x, \ldots),$$

where the two terms with a repeated argument have been dropped. For the converse, if exchanging two arguments changes the sign, then a value with the same vector x in two positions equals its own negative, so twice it is 0, and it is 0 when $2 \neq 0$ in k. □

9.10.2 Symmetric and Exterior Powers as Quotients

Let $J_d(V) \subseteq V^{\otimes d}$ be the subspace spanned by the tensors $\eta - \sigma\cdot\eta$ for $\eta \in V^{\otimes d}$ and $\sigma \in S_d$, and let $I_d(V) \subseteq V^{\otimes d}$ be the subspace spanned by the pure tensors $v_1 \otimes \cdots \otimes v_d$ with $v_i = v_j$ for some $i \neq j$. For $d \leqslant 1$ both are zero.

Definition 9.10.3 (Symmetric power)

The dth ***symmetric power*** of V is the quotient space

$$\operatorname{Sym}^d V = V^{\otimes d}/J_d(V).$$

The class of $v_1 \otimes \cdots \otimes v_d$ is written $v_1 v_2 \cdots v_d$.

Definition 9.10.4 (Exterior power)

The dth ***exterior power*** of V is the quotient space

$$\bigwedge^d V = V^{\otimes d}/I_d(V).$$

The class of $v_1 \otimes \cdots \otimes v_d$ is written $v_1 \wedge \cdots \wedge v_d$ and called a ***wedge product***.

In $\operatorname{Sym}^d V$ the factors of a product $v_1 \cdots v_d$ may be permuted freely, since $v_1 \otimes \cdots \otimes v_d - \sigma \cdot (v_1 \otimes \cdots \otimes v_d)$ lies in $J_d(V)$. In $\bigwedge^d V$ a wedge product with a repeated factor is zero. Since

$$v \otimes w + w \otimes v = (v+w) \otimes (v+w) - v \otimes v - w \otimes w \in I_2(V),$$

we get $v \wedge w = -w \wedge v$; the same computation in any two positions shows that exchanging two factors of a wedge product changes its sign. For small d, $\operatorname{Sym}^0 V = \bigwedge^0 V = k$ and $\operatorname{Sym}^1 V = \bigwedge^1 V = V$.

Proposition 9.10.5 (Universal properties)

(i) The map $V^d \to \operatorname{Sym}^d V$, $(v_1, \ldots, v_d) \mapsto v_1 \cdots v_d$, is multilinear and symmetric. For every symmetric multilinear map $m : V^d \to U$ there is a unique linear map $\tilde{m} : \operatorname{Sym}^d V \to U$ with $\tilde{m}(v_1 \cdots v_d) = m(v_1, \ldots, v_d)$.

(ii) The map $V^d \to \bigwedge^d V$, $(v_1, \ldots, v_d) \mapsto v_1 \wedge \cdots \wedge v_d$, is multilinear and alternating. For every alternating multilinear map $m : V^d \to U$ there is a unique linear map $\tilde{m} : \bigwedge^d V \to U$ with $\tilde{m}(v_1 \wedge \cdots \wedge v_d) = m(v_1, \ldots, v_d)$.

Consequently $\operatorname{Hom}(\operatorname{Sym}^d V, U)$ is isomorphic to the space of symmetric multilinear maps $V^d \to U$, and $\operatorname{Hom}(\bigwedge^d V, U)$ to the space of alternating multilinear maps $V^d \to U$.

Proof. The two maps are the composites of $\mu : V^d \to V^{\otimes d}$ with the quotient maps, so they are multilinear. The first is symmetric because $v_{\sigma(1)} \otimes \cdots \otimes v_{\sigma(d)} =$

$\sigma^{-1}\cdot(v_1\otimes\cdots\otimes v_d)$ has the same class modulo $J_d(V)$ as $v_1\otimes\cdots\otimes v_d$; the second is alternating by the definition of $I_d(V)$.

(i) Given m, let $\hat{m}: V^{\otimes d}\to U$ be the linear map with $\hat{m}(v_1\otimes\cdots\otimes v_d)=m(v_1,\dots,v_d)$ (Theorem 9.9.2). For each σ, the linear map $\eta\mapsto\hat{m}(\eta)-\hat{m}(\sigma\cdot\eta)$ vanishes on pure tensors because m is symmetric, so it vanishes identically. Hence $J_d(V)\subseteq\ker\hat{m}$, and $\hat{m}$ factors through the quotient $\operatorname{Sym}^d V$.

(ii) If m is alternating, $\hat{m}$ vanishes on the spanning elements of $I_d(V)$, so it factors through $\bigwedge^d V$.

In both cases $\tilde{m}$ is unique because the classes of pure tensors span the quotient. The last statement follows as for $V\otimes W$: $m\mapsto\tilde{m}$ is a linear bijection, with inverse given by composing with the map $V^d\to\operatorname{Sym}^d V$ or $V^d\to\bigwedge^d V$. □

Proposition 9.10.6 (Induced maps)

Let $\varphi: V\to W$ be a linear map between finite-dimensional spaces. There are unique linear maps $\operatorname{Sym}^d\varphi:\operatorname{Sym}^d V\to\operatorname{Sym}^d W$ and $\bigwedge^d\varphi:\bigwedge^d V\to\bigwedge^d W$ with

$$\operatorname{Sym}^d\varphi\,(v_1\cdots v_d)=\varphi(v_1)\cdots\varphi(v_d),$$
$$\bigwedge^d\varphi\,(v_1\wedge\cdots\wedge v_d)=\varphi(v_1)\wedge\cdots\wedge\varphi(v_d).$$

They satisfy $\bigwedge^d(\psi\circ\varphi)=\bigwedge^d\psi\circ\bigwedge^d\varphi$ and $\bigwedge^d(\mathrm{id}_V)=\mathrm{id}$, and likewise for Sym^d.

Proof. The map $(v_1,\dots,v_d)\mapsto\varphi(v_1)\wedge\cdots\wedge\varphi(v_d)$ is multilinear, and alternating because $v_i=v_j$ implies $\varphi(v_i)=\varphi(v_j)$. Proposition 9.10.5 gives $\bigwedge^d\varphi$. Both sides of each identity are linear and agree on wedge products, which span. The case of Sym^d is the same. □

These induced maps are what make it possible to form the symmetric and exterior powers of a representation, a notion defined in Chapter 13 (Definition 13.1.1).

9.10.3 Bases and Dimensions

Fix a basis $e_1,\dots,e_n$ of V. For a d-tuple $J=(j_1,\dots,j_d)$ of indices in $\{1,\dots,n\}$, write

$$t_J=e_{j_1}\otimes\cdots\otimes e_{j_d};$$

by Theorem 9.9.2, these n^d tensors form a basis of $V^{\otimes d}$. For $\sigma\in S_d$ we have $\sigma\cdot t_J=t_{\sigma J}$, where $\sigma J=(j_{\sigma^{-1}(1)},\dots,j_{\sigma^{-1}(d)})$ is J with its entries rearranged.

Proposition 9.10.7 (Basis of $\operatorname{Sym}^d V$)
For $\alpha = (\alpha_1, \dots, \alpha_n)$ with integers $\alpha_i \geqslant 0$ and $\alpha_1 + \cdots + \alpha_n = d$, let

$$e^\alpha = e_1^{\alpha_1} e_2^{\alpha_2} \cdots e_n^{\alpha_n} \in \operatorname{Sym}^d V$$

be the product of α_1 copies of e_1, α_2 copies of e_2, and so on. These monomials form a basis of $\operatorname{Sym}^d V$, and

$$\dim \operatorname{Sym}^d V = \binom{n+d-1}{d}.$$

Proof. Spanning. The classes of the t_J span $\operatorname{Sym}^d V$. The class of t_J is $e_{j_1} \cdots e_{j_d}$; since its factors may be permuted, it equals $e^{\alpha(J)}$, where $\alpha(J)_i$ is the number of entries of J equal to i.

Independence. Let E be a vector space with a basis (ε_α) indexed by the exponent vectors α of total d, and let $c : V^{\otimes d} \to E$ be the linear map with $c(t_J) = \varepsilon_{\alpha(J)}$. Rearranging the entries of J does not change $\alpha(J)$, so $c(t_J - \sigma \cdot t_J) = 0$ for all J and σ. Since the t_J span $V^{\otimes d}$ and $\eta \mapsto \eta - \sigma \cdot \eta$ is linear, c vanishes on $J_d(V)$, so it induces a linear map $\bar{c} : \operatorname{Sym}^d V \to E$ with $\bar{c}(e^\alpha) = \varepsilon_\alpha$. The ε_α are linearly independent, hence so are the e^α.

Counting. Exponent vectors correspond to arrangements of d stars and $n-1$ bars in a row (α_i is the number of stars between the $(i-1)$st and the ith bar), and there are $\binom{n+d-1}{d}$ such arrangements. □

Proposition 9.10.8 (Basis of $\bigwedge^d V$)
For an increasing tuple $I = (i_1 < i_2 < \cdots < i_d)$, write $e_I = e_{i_1} \wedge e_{i_2} \wedge \cdots \wedge e_{i_d}$. These $\binom{n}{d}$ elements form a basis of $\bigwedge^d V$, so

$$\dim \bigwedge^d V = \binom{n}{d}.$$

In particular $\bigwedge^d V = 0$ for $d > n$.

Proof. Let $q : V^{\otimes d} \to \bigwedge^d V$ be the quotient map. Lemma 9.10.2(ii), applied to the alternating map $(v_1, \dots, v_d) \mapsto v_1 \wedge \cdots \wedge v_d$, gives

$$q\big(\sigma \cdot (v_1 \otimes \cdots \otimes v_d)\big) = v_{\sigma^{-1}(1)} \wedge \cdots \wedge v_{\sigma^{-1}(d)} = \operatorname{sgn}(\sigma^{-1})\, v_1 \wedge \cdots \wedge v_d,$$

and $\operatorname{sgn}(\sigma^{-1}) = \operatorname{sgn}(\sigma)$; by linearity, $q(\sigma \cdot \eta) = \operatorname{sgn}(\sigma)\, q(\eta)$ for all $\eta \in V^{\otimes d}$.

Spanning. The classes $q(t_J)$ span $\bigwedge^d V$. If J has a repeated entry, $q(t_J) = 0$. If the entries of J are distinct, let I be the increasing tuple with the same entries; there

is a unique $\rho_J \in S_d$ with $J = \rho_J I$, that is, $t_J = \rho_J \cdot t_I$, and then $q(t_J) = \operatorname{sgn}(\rho_J)\, e_I$. Hence the e_I span.

Independence. Let E be a vector space with a basis (ε_I) indexed by the increasing d-tuples, and let $c : V^{\otimes d} \to E$ be the linear map with $c(t_J) = 0$ if J has a repeated entry and $c(t_J) = \operatorname{sgn}(\rho_J)\, \varepsilon_I$ otherwise (with I and ρ_J as above). We show that c vanishes on $I_d(V)$. Let $x = v_1 \otimes \cdots \otimes v_d$ with $v_a = v_b$ for some $a < b$, let $\tau = (a\ b)$, and write $v_l = \sum_j x_{lj} e_j$. Expanding,

$$x = \sum_J x_J\, t_J, \qquad x_J = x_{1j_1} x_{2j_2} \cdots x_{dj_d},$$

and $x_{\tau J} = x_J$ because $v_a = v_b$ (the tuple τJ is J with its ath and bth entries exchanged). The terms with $j_a = j_b$ have $c(t_J) = 0$. The remaining tuples come in pairs $\{J, \tau J\}$ with $J \neq \tau J$, and $c(t_{\tau J}) = -c(t_J)$: either both tuples have a repeated entry, or $t_{\tau J} = \tau \cdot t_J = (\tau\rho_J) \cdot t_I$, so that $\rho_{\tau J} = \tau\rho_J$ and $\operatorname{sgn}(\rho_{\tau J}) = -\operatorname{sgn}(\rho_J)$. So the contributions of each pair cancel, and $c(x) = 0$. Thus $I_d(V) \subseteq \ker c$, and c induces a linear map $\bar{c} : \bigwedge^d V \to E$ with $\bar{c}(e_I) = c(t_I) = \varepsilon_I$. The ε_I are independent, hence so are the e_I.

If $d > n$ there are no increasing d-tuples, so $\bigwedge^d V = 0$. □

Corollary 9.10.9 (The top exterior power)

If $\dim V = n$, then $\bigwedge^n V$ is one-dimensional, and for every basis $e_1, \ldots, e_n$ of V the element $e_1 \wedge \cdots \wedge e_n$ is a basis of $\bigwedge^n V$.

The line $\bigwedge^n V$ is called the ***determinant line*** of V; Theorem 9.11.3 below explains the name.

NOTE Compare the three dimension formulas, for $\dim V = n$:

- $\dim V^{\otimes d} = n^d$, the number of ordered d-tuples of indices;
- $\dim \operatorname{Sym}^d V = \binom{n+d-1}{d}$, the number of multisets of size d from n indices;
- $\dim \bigwedge^d V = \binom{n}{d}$, the number of subsets of size d of n indices.

Tensor powers remember the order of the factors; symmetric powers allow repeated factors and forget their order; exterior powers forbid repeated factors and remember their order only up to sign.

Proposition 9.10.10 (Wedge products detect linear independence)

Vectors $v_1, \ldots, v_d \in V$ are linearly independent if and only if $v_1 \wedge \cdots \wedge v_d \neq 0$ in $\bigwedge^d V$.

Proof. If the v_i are dependent, $v_1 \wedge \cdots \wedge v_d = 0$ by Lemma 9.10.2(iii), applied to the alternating map $(v_1, \ldots, v_d) \mapsto v_1 \wedge \cdots \wedge v_d$. If they are independent, extend them to a basis $v_1, \ldots, v_n$ of V. Then $v_1 \wedge \cdots \wedge v_d$ is the basis element e_I, $I = (1, \ldots, d)$, of Proposition 9.10.8 for this basis, so it is nonzero. □

9.10.4 Symmetric and Alternating Tensors

Many books define $\operatorname{Sym}^d V$ and $\bigwedge^d V$ as subspaces of $V^{\otimes d}$. When $d!$ is invertible in k, the subspaces are isomorphic to our quotients.

Definition 9.10.11 (Symmetric and alternating tensors)

A tensor $\eta \in V^{\otimes d}$ is ***symmetric*** if $\sigma \cdot \eta = \eta$ for all $\sigma \in S_d$, and ***alternating*** if $\sigma \cdot \eta = \operatorname{sgn}(\sigma)\,\eta$ for all $\sigma \in S_d$. They form subspaces $\mathcal{S}^d(V)$ and $\mathcal{A}^d(V)$ of $V^{\otimes d}$.

Proposition 9.10.12 (Tensor models when $d!$ is invertible)

Suppose $d! \neq 0$ in k (for instance, $\operatorname{char} k = 0$). Define linear maps $P_{\mathcal{S}}, P_{\mathcal{A}} : V^{\otimes d} \to V^{\otimes d}$ by

$$P_{\mathcal{S}}(\eta) = \frac{1}{d!} \sum_{\sigma \in S_d} \sigma \cdot \eta, \qquad P_{\mathcal{A}}(\eta) = \frac{1}{d!} \sum_{\sigma \in S_d} \operatorname{sgn}(\sigma)\, \sigma \cdot \eta.$$

Then $P_{\mathcal{S}}$ is a projection onto $\mathcal{S}^d(V)$ with kernel $J_d(V)$, and $P_{\mathcal{A}}$ is a projection onto $\mathcal{A}^d(V)$ with kernel $I_d(V)$. Consequently

$$V^{\otimes d} = \mathcal{S}^d(V) \oplus J_d(V) = \mathcal{A}^d(V) \oplus I_d(V),$$

and the quotient maps restrict to isomorphisms $\mathcal{S}^d(V) \to \operatorname{Sym}^d V$ and $\mathcal{A}^d(V) \to \bigwedge^d V$, whose inverses are

$$v_1 \cdots v_d \mapsto \frac{1}{d!} \sum_{\sigma \in S_d} v_{\sigma(1)} \otimes \cdots \otimes v_{\sigma(d)},$$

$$v_1 \wedge \cdots \wedge v_d \mapsto \frac{1}{d!} \sum_{\sigma \in S_d} \operatorname{sgn}(\sigma)\, v_{\sigma(1)} \otimes \cdots \otimes v_{\sigma(d)}.$$

Proof. We treat $P_{\mathcal{A}}$; the argument for $P_{\mathcal{S}}$ is the same with the signs omitted. For $\tau \in S_d$, substituting $\rho = \tau\sigma$ and using $\operatorname{sgn}(\sigma) = \operatorname{sgn}(\tau)\operatorname{sgn}(\rho)$,

$$\tau \cdot P_{\mathcal{A}}(\eta) = \frac{1}{d!} \sum_{\sigma} \operatorname{sgn}(\sigma)\, (\tau\sigma) \cdot \eta = \operatorname{sgn}(\tau)\, P_{\mathcal{A}}(\eta),$$

so the image of $P_{\mathcal{A}}$ lies in $\mathcal{A}^d(V)$. If $\eta \in \mathcal{A}^d(V)$, each term $\operatorname{sgn}(\sigma)\,\sigma\cdot\eta$ equals $\operatorname{sgn}(\sigma)^2\eta = \eta$, so $P_{\mathcal{A}}(\eta) = \eta$. Hence $P_{\mathcal{A}}$ is a projection onto $\mathcal{A}^d(V)$, and $V^{\otimes d} = \mathcal{A}^d(V) \oplus \ker P_{\mathcal{A}}$.

Let $q : V^{\otimes d} \to \bigwedge^d V$ be the quotient map. As in the proof of Proposition 9.10.8, $q(\sigma\cdot\eta) = \operatorname{sgn}(\sigma)\,q(\eta)$, so $q(P_{\mathcal{A}}(\eta)) = \frac{1}{d!}\sum_\sigma \operatorname{sgn}(\sigma)^2 q(\eta) = q(\eta)$; that is, $\eta - P_{\mathcal{A}}(\eta) \in I_d(V)$ for every η. If $P_{\mathcal{A}}(\eta) = 0$, this says $\eta \in I_d(V)$. Conversely, let $x = v_1\otimes\cdots\otimes v_d$ with $v_a = v_b$ for some $a \neq b$, and let $\tau = (a\ b)$. Then $\tau\cdot x = x$, and substituting $\rho = \sigma\tau$ gives $P_{\mathcal{A}}(\tau\cdot x) = \operatorname{sgn}(\tau)P_{\mathcal{A}}(x)$, so $P_{\mathcal{A}}(x) = -P_{\mathcal{A}}(x)$. As $d \geqslant 2$, 2 divides $d!$, so $2 \neq 0$ in k and $P_{\mathcal{A}}(x) = 0$. Hence $\ker P_{\mathcal{A}} = I_d(V) = \ker q$. It follows that q restricts to an isomorphism $\mathcal{A}^d(V) \to \bigwedge^d V$, with inverse $q(\eta) \mapsto P_{\mathcal{A}}(\eta)$. For $\eta = v_1\otimes\cdots\otimes v_d$ this inverse gives $\frac{1}{d!}\sum_\sigma \operatorname{sgn}(\sigma)\, v_{\sigma^{-1}(1)}\otimes\cdots\otimes v_{\sigma^{-1}(d)}$, which is the stated formula after replacing σ by σ^{-1}.

For $P_{\mathcal{S}}$, the identity $\eta - P_{\mathcal{S}}(\eta) = \frac{1}{d!}\sum_\sigma(\eta - \sigma\cdot\eta)$ shows $\ker P_{\mathcal{S}} \subseteq J_d(V)$, and $P_{\mathcal{S}}(\sigma\cdot\eta) = P_{\mathcal{S}}(\eta)$ shows $J_d(V) \subseteq \ker P_{\mathcal{S}}$. □

NOTE Under these isomorphisms $e_1 \wedge e_2 \in \bigwedge^2 V$ corresponds to the alternating tensor $\frac{1}{2}(e_1\otimes e_2 - e_2\otimes e_1)$. Some authors identify $v_1\wedge\cdots\wedge v_d$ with $\sum_\sigma \operatorname{sgn}(\sigma)\, v_{\sigma(1)}\otimes\cdots\otimes v_{\sigma(d)}$, without the factor $\frac{1}{d!}$; this rescales the identification by $d!$. When $d! = 0$ in k, the subspace models break down: Exercise 9.11.14(1) shows that over a field of characteristic 2, $\mathcal{A}^2(k^2)$ is three-dimensional, while $\bigwedge^2 k^2$ is one-dimensional. This is why we define $\operatorname{Sym}^d V$ and $\bigwedge^d V$ as quotients.

Corollary 9.10.13 (Symmetric and alternating parts of $V \otimes V$)

Suppose $\operatorname{char} k \neq 2$, and let $\tau : V\otimes V \to V\otimes V$ be the swap, $\tau(v\otimes w) = w\otimes v$. Then

$$V \otimes V = \mathcal{S}^2(V) \oplus \mathcal{A}^2(V) \cong \operatorname{Sym}^2 V \oplus \bigwedge^2 V,$$

where τ acts as $+1$ on $\mathcal{S}^2(V)$ and as -1 on $\mathcal{A}^2(V)$. Under the isomorphism $V^*\otimes V^* \cong B(V)$ of Proposition 9.8.13(ii), $\mathcal{S}^2(V^*)$ corresponds to the symmetric bilinear forms and $\mathcal{A}^2(V^*)$ to the skew-symmetric ones. Hence $B(V)$ is the direct sum of the symmetric and the skew-symmetric forms, and

$$(V \otimes V)^* \cong \operatorname{Sym}^2(V^*) \oplus \bigwedge^2(V^*).$$

Proof. The swap is the action of the transposition in S_2, so $\mathcal{S}^2(V) = \ker(\tau - \mathrm{id})$ and $\mathcal{A}^2(V) = \ker(\tau + \mathrm{id})$. Since $\tau^2 = \mathrm{id}$, every η is the sum $\frac{1}{2}(\eta + \tau\eta) + \frac{1}{2}(\eta - \tau\eta)$ of an element of $\mathcal{S}^2(V)$ and an element of $\mathcal{A}^2(V)$; and if η lies in both, then $\eta = \tau\eta = -\eta$, so $\eta = 0$. This is also the case $d = 2$ of Proposition 9.10.12, which gives the isomorphisms with $\operatorname{Sym}^2 V$ and $\bigwedge^2 V$. Under Proposition 9.8.13(ii), $\ell\otimes m$

corresponds to $b(v,w) = \ell(v)m(w)$ and $\tau(\ell\otimes m) = m\otimes\ell$ to $m(v)\ell(w) = b(w,v)$; by linearity, τ corresponds to $b \mapsto b'$ with $b'(v,w) = b(w,v)$, whose ±1-eigenspaces are the symmetric and the skew-symmetric forms. Finally $(V\otimes V)^* \cong V^*\otimes V^*$ by Proposition 9.8.13(i), and we apply the first statement to V^*. □

9.10.5 The Symmetric Algebra and Polynomials

The product of the tensor algebra descends to symmetric powers. For $a, b \geqslant 0$, the map $V^a \times V^b \to \operatorname{Sym}^{a+b} V$ given by

$$(v_1,\dots,v_a,w_1,\dots,w_b) \mapsto v_1\cdots v_a w_1\cdots w_b$$

is multilinear and symmetric in the v_i for fixed w_j, so by Proposition 9.10.5 it induces, for each fixed $(w_1,\dots,w_b)$, a linear map $\operatorname{Sym}^a V \to \operatorname{Sym}^{a+b} V$. For a fixed $\eta \in \operatorname{Sym}^a V$, the image of η is multilinear and symmetric in $(w_1,\dots,w_b)$: this holds when $\eta = v_1\cdots v_a$, and such elements span $\operatorname{Sym}^a V$. So it induces a linear map $\operatorname{Sym}^b V \to \operatorname{Sym}^{a+b} V$. The result is a bilinear product $\operatorname{Sym}^a V \times \operatorname{Sym}^b V \to \operatorname{Sym}^{a+b} V$ with

$$(v_1\cdots v_a)(w_1\cdots w_b) = v_1\cdots v_a w_1\cdots w_b.$$

Definition 9.10.14 (Symmetric algebra)

The ***symmetric algebra*** of V is

$$\operatorname{Sym}^\bullet V = \bigoplus_{d=0}^{\infty} \operatorname{Sym}^d V,$$

with the product that extends the products $\operatorname{Sym}^a V \times \operatorname{Sym}^b V \to \operatorname{Sym}^{a+b} V$ bilinearly.

The ***polynomial ring*** $k[x_1,\dots,x_n]$ in n variables is the vector space with basis the monomials $x^a = x_1^{a_1}\cdots x_n^{a_n}$, $a \in \mathbb{Z}^n_{\geqslant 0}$, with the multiplication $x^a x^b = x^{a+b}$ extended bilinearly; for $n = 1$ it is the ring $k[x]$. The ***homogeneous polynomials of degree*** d are the linear combinations of the x^a with $a_1 + \cdots + a_n = d$.

Theorem 9.10.15 (The symmetric algebra is a polynomial ring)

$\operatorname{Sym}^\bullet V$ is a commutative ring with identity $1 \in \operatorname{Sym}^0 V = k$, and its multiplication is k-bilinear. If $e_1,\dots,e_n$ is a basis of V, the linear map

$$\operatorname{Sym}^\bullet V \to k[x_1,\dots,x_n], \qquad e^a \mapsto x^a,$$

is an isomorphism of rings. In particular, $\mathrm{Sym}^d V$ is isomorphic to the space of homogeneous polynomials of degree d in n variables.

Proof. The product is bilinear by construction, so it distributes over addition. Associativity can be checked on products of vectors, where both sides are $v_1 \cdots v_a w_1 \cdots w_b u_1 \cdots u_c$; commutativity too, since $w_1 \cdots w_b v_1 \cdots v_a$ differs from $v_1 \cdots v_a w_1 \cdots w_b$ only by a permutation of the factors. The element $1 \in \mathrm{Sym}^0 V$ is an identity. By Proposition 9.10.7, the monomials e^a (of all degrees) form a basis of $\mathrm{Sym}^\bullet V$, so the map $e^a \mapsto x^a$ is a linear isomorphism. It is multiplicative on basis elements, because $e^a e^b = e^{a+b}$ (rearrange the factors) and $x^a x^b = x^{a+b}$; both products are bilinear, so the map is multiplicative. □

NOTE $\mathrm{Sym}^\bullet V$ is the coordinate-free version of the polynomial ring: it depends only on V, and a choice of basis identifies it with $k[x_1, \ldots, x_n]$.

The most natural instance is V^*. If $e_1, \ldots, e_n$ is a basis of V and $x_i = e_i^*$, then $x_i(v)$ is the ith coordinate of v, and a product of the x_i should be a polynomial function of the coordinates. A function $f : V \to k$ is a ***homogeneous polynomial function of degree*** d if $f(\sum_i c_i e_i) = p(c_1, \ldots, c_n)$ for a homogeneous polynomial p of degree d; a linear change of coordinates turns homogeneous polynomials of degree d into homogeneous polynomials of degree d, so this does not depend on the basis.

Proposition 9.10.16 ($\mathrm{Sym}^d(V^*)$ and polynomial functions)

Let $e_1, \ldots, e_n$ be a basis of V and $x_i = e_i^*$.

(i) There is a unique linear map P from $\mathrm{Sym}^d(V^*)$ to the space of functions $V \to k$ with $P(\ell_1 \cdots \ell_d)(v) = \ell_1(v) \cdots \ell_d(v)$.

(ii) P sends the monomial $x^a \in \mathrm{Sym}^d(V^*)$ to the function $v \mapsto x_1(v)^{a_1} \cdots x_n(v)^{a_n}$. So the image of P is the space of homogeneous polynomial functions of degree d on V.

(iii) If k is infinite, P is injective, and so is the map it induces on $\mathrm{Sym}^\bullet(V^*) \cong k[x_1, \ldots, x_n]$. Thus $\mathrm{Sym}^d(V^*)$ is identified with the homogeneous polynomial functions of degree d on V, and $\mathrm{Sym}^\bullet(V^*)$ with the ring of polynomial functions on V, the variable x_i being the ith coordinate function.

Proof. (i) The map $(\ell_1, \ldots, \ell_d) \mapsto (v \mapsto \ell_1(v) \cdots \ell_d(v))$ is multilinear and symmetric, with values in the vector space of functions $V \to k$; apply Proposition 9.10.5. (ii) follows from (i).

(iii) Extend P to $\mathrm{Sym}^\bullet(V^*)$ degree by degree. By Theorem 9.10.15, an element of $\mathrm{Sym}^\bullet(V^*)$ is a polynomial $p = \sum_a c_a x^a$, and P sends it to the function $v \mapsto p(x_1(v), \dots, x_n(v))$. So it suffices to show that a polynomial $p \in k[x_1, \dots, x_n]$ that vanishes at every point of k^n is zero. We use induction on n. For $n = 1$, a nonzero polynomial has at most $\deg p$ roots (Corollary 4.3.15), while k is infinite. For $n > 1$, write $p = \sum_j p_j(x_1, \dots, x_{n-1})\, x_n^j$. For each point $c \in k^{n-1}$, the one-variable polynomial $\sum_j p_j(c)\, x_n^j$ vanishes on all of k, so every $p_j(c) = 0$. Thus each p_j vanishes on k^{n-1}, and $p_j = 0$ by induction. □

NOTE When k is finite, P need not be injective. Over $\mathbb{F}_2$, the monomials $x_1^2 x_2$ and $x_1 x_2^2$ in $\mathrm{Sym}^3(V^*)$ define the same function, because $c^2 = c$ for every $c \in \mathbb{F}_2$. So over finite fields $\mathrm{Sym}^d(V^*)$ consists of polynomials as algebraic expressions, and these carry more information than the functions they define.

NOTE When $d! \neq 0$ in k, Propositions 9.10.12 and 9.9.6 identify $\mathrm{Sym}^d(V^*)$ with the symmetric tensors in $(V^*)^{\otimes d}$, that is, with the symmetric d-linear forms on V. Following $\ell_1 \cdots \ell_d$ through these identifications gives the form

$$\eta(v_1, \dots, v_d) = \frac{1}{d!} \sum_{\sigma \in S_d} \ell_{\sigma(1)}(v_1) \cdots \ell_{\sigma(d)}(v_d),$$

and $\eta(v, \dots, v) = \ell_1(v) \cdots \ell_d(v)$. So the polynomial function attached to a symmetric multilinear form η is $v \mapsto \eta(v, \dots, v)$. For $d = 2$ this is the passage from a symmetric bilinear form to its quadratic form (Section 9.6), and polarization (Proposition 9.6.2) recovers the form from the function.

9.10.6 The Exterior Algebra

Exactly as for symmetric powers, the map $(v_1, \dots, v_a, w_1, \dots, w_b) \mapsto v_1 \wedge \cdots \wedge v_a \wedge w_1 \wedge \cdots \wedge w_b$ is multilinear and alternating in the v_i for fixed w_j, and in the w_j for fixed v_i. Applying Proposition 9.10.5 once in each variable gives a bilinear product $\bigwedge^a V \times \bigwedge^b V \to \bigwedge^{a+b} V$ with

$$(v_1 \wedge \cdots \wedge v_a) \wedge (w_1 \wedge \cdots \wedge w_b) = v_1 \wedge \cdots \wedge v_a \wedge w_1 \wedge \cdots \wedge w_b.$$

Definition 9.10.17 (Exterior algebra)

The ***exterior algebra*** (or ***Grassmann algebra***) of V is

$$\bigwedge^\bullet V = \bigoplus_{d=0}^{n} \bigwedge^d V,$$

with the product $\wedge$ that extends the products $\bigwedge^a V \times \bigwedge^b V \to \bigwedge^{a+b} V$ bilinearly. (Since $\bigwedge^d V = 0$ for $d > n$, the sum stops at n.)

Proposition 9.10.18 (Properties of the exterior algebra)

Let $\dim V = n$.

(i) $\bigwedge^\bullet V$ is a ring with identity $1 \in \bigwedge^0 V = k$, its multiplication is k-bilinear, and $\dim \bigwedge^\bullet V = 2^n$.

(ii) The product is ***graded-commutative***: for $\eta \in \bigwedge^a V$ and $\theta \in \bigwedge^b V$,

$$\theta \wedge \eta = (-1)^{ab}\, \eta \wedge \theta.$$

Proof. (i) Distributivity holds because the product is bilinear, and associativity and the identity can be checked on wedge products of vectors, where the product is concatenation. By Proposition 9.10.8, $\dim \bigwedge^\bullet V = \sum_{d=0}^{n} \binom{n}{d} = (1+1)^n = 2^n$.

(ii) Both sides are bilinear in (η, θ), so it suffices to take $\eta = v_1 \wedge \cdots \wedge v_a$ and $\theta = w_1 \wedge \cdots \wedge w_b$. Starting from $v_1 \wedge \cdots \wedge v_a \wedge w_1 \wedge \cdots \wedge w_b$, move v_a to the right past the b factors $w_1, \ldots, w_b$, then v_{a-1}, and so on. This reaches $w_1 \wedge \cdots \wedge w_b \wedge v_1 \wedge \cdots \wedge v_a$ after ab exchanges of adjacent factors, each of which changes the sign (Lemma 9.10.2(i)). □

9.10.7 Operators on Tensor, Symmetric, and Exterior Powers

For a diagonalizable operator, the induced operators on tensor, symmetric and exterior powers are easy to describe; this is what Chapter 13 uses to compute the characters of $\mathrm{Sym}^2 V$ and $\bigwedge^2 V$ (Definition 13.4.1).

Proposition 9.10.19 (Eigenvalues on tensor, symmetric, and exterior powers)

Let $\varphi \in \mathrm{End}(V)$, and suppose $e_1, \ldots, e_n$ is a basis of V with $\varphi(e_i) = \lambda_i e_i$. With the notation of Propositions 9.10.7 and 9.10.8:

(i) $\varphi^{\otimes d}(t_J) = \lambda_{j_1} \cdots \lambda_{j_d}\, t_J$;

(ii) $\mathrm{Sym}^d \varphi\,(e^a) = \lambda_1^{a_1} \cdots \lambda_n^{a_n}\, e^a$;

(iii) $\bigwedge^d \varphi\,(e_I) = \lambda_{i_1}\cdots\lambda_{i_d}\,e_I$.

In particular $\varphi^{\otimes d}$, $\operatorname{Sym}^d\varphi$ and $\bigwedge^d\varphi$ are diagonalizable, and

$$\operatorname{Tr}(\varphi^{\otimes d}) = (\operatorname{Tr}\varphi)^d,$$

$$\operatorname{Tr}\left(\operatorname{Sym}^d\varphi\right) = \sum_{i_1\leqslant\cdots\leqslant i_d} \lambda_{i_1}\cdots\lambda_{i_d}, \qquad \operatorname{Tr}\left(\bigwedge^d\varphi\right) = \sum_{i_1<\cdots<i_d} \lambda_{i_1}\cdots\lambda_{i_d}.$$

If $\operatorname{char} k \neq 2$, then

$$\operatorname{Tr}\left(\operatorname{Sym}^2\varphi\right) = \tfrac{1}{2}\big((\operatorname{Tr}\varphi)^2 + \operatorname{Tr}(\varphi^2)\big),$$
$$\operatorname{Tr}\left(\bigwedge^2\varphi\right) = \tfrac{1}{2}\big((\operatorname{Tr}\varphi)^2 - \operatorname{Tr}(\varphi^2)\big).$$

Proof. By the definitions of the induced maps, $\bigwedge^d\varphi\,(e_{i_1}\wedge\cdots\wedge e_{i_d}) = \lambda_{i_1}e_{i_1}\wedge\cdots\wedge\lambda_{i_d}e_{i_d} = \lambda_{i_1}\cdots\lambda_{i_d}\,e_I$, and (i) and (ii) are the same computation. So each of the three bases consists of eigenvectors, and each trace is the sum of the eigenvalues. The monomials e^a correspond to the nondecreasing tuples $i_1\leqslant\cdots\leqslant i_d$ (list each index i exactly a_i times), which gives the formula for $\operatorname{Tr}(\operatorname{Sym}^d\varphi)$; the sum over all d-tuples J of $\lambda_{j_1}\cdots\lambda_{j_d}$ is $(\sum_i\lambda_i)^d$. For $d=2$, $\operatorname{Tr}\varphi = \sum_i\lambda_i$ and $\operatorname{Tr}(\varphi^2) = \sum_i\lambda_i^2$, since $\varphi^2(e_i) = \lambda_i^2 e_i$. From $(\sum_i\lambda_i)^2 = \sum_i\lambda_i^2 + 2\sum_{i<j}\lambda_i\lambda_j$ we get $\operatorname{Tr}(\bigwedge^2\varphi) = \sum_{i<j}\lambda_i\lambda_j = \frac{1}{2}((\operatorname{Tr}\varphi)^2 - \operatorname{Tr}(\varphi^2))$, and $\operatorname{Tr}(\operatorname{Sym}^2\varphi) = \sum_{i<j}\lambda_i\lambda_j + \sum_i\lambda_i^2 = \frac{1}{2}((\operatorname{Tr}\varphi)^2 + \operatorname{Tr}(\varphi^2))$. □

9.11 Determinants, Trace, and Volume Forms

Two of the most important invariants of an operator φ on a finite-dimensional vector space, its trace and its determinant, were defined through matrices (Definitions 5.3.11 and 5.4.20), and their independence of the basis needed a computation. Tensor products and exterior powers produce both of them without choosing a basis. Throughout this section V is finite-dimensional.

9.11.1 Trace as Contraction

Under the isomorphism $\operatorname{End}(V) = \operatorname{Hom}(V,V) \cong V^*\otimes V$ of Proposition 9.8.10, the trace becomes a natural pairing.

Definition 9.11.1 (Contraction)

The ***contraction*** is the linear functional

$$c : V^*\otimes V \to k, \qquad c(\ell\otimes v) = \ell(v);$$

it exists by the universal property, since $(\ell, v) \mapsto \ell(v)$ is bilinear.

Proposition 9.11.2 (Trace as contraction)
Let $\varphi \in \mathrm{End}(V)$, and let $\eta_\varphi \in V^* \otimes V$ be the tensor corresponding to φ under Proposition 9.8.10. Then

$$\mathrm{Tr}(\varphi) = c(\eta_\varphi).$$

In particular, the trace of the rank-one map $v \mapsto \ell(v)\, w$ is $\ell(w)$.

Proof. Choose a basis $e_1, \dots, e_n$ of V, with dual basis (e_i^*), and let $A = (a_{ij})$ be the matrix of φ, so $\varphi(e_j) = \sum_i a_{ij} e_i$. The tensor $e_j^* \otimes e_i$ corresponds to the map sending e_j to e_i and the other basis vectors to 0, whose matrix has a single 1, in row i and column j. Hence

$$\eta_\varphi = \sum_{i,j} a_{ij}\, e_j^* \otimes e_i, \qquad c(\eta_\varphi) = \sum_{i,j} a_{ij}\, e_j^*(e_i) = \sum_{i,j} a_{ij}\, \delta_{ij} = \sum_i a_{ii} = \mathrm{Tr}(\varphi).$$

For the rank-one map, $\eta = \ell \otimes w$ and $c(\eta) = \ell(w)$. □

NOTE Since c involves no basis, this gives a basis-free definition of the trace: it is the composite $\mathrm{End}(V) \cong V^* \otimes V \xrightarrow{c} k$. The formula $\mathrm{Tr}(A) = \sum_i a_{ii}$ is what the contraction looks like in coordinates, and the invariance of the trace under change of basis becomes automatic.

9.11.2 The Determinant via the Top Exterior Power

Let $\dim V = n \geqslant 1$. An operator φ on V induces an operator $\bigwedge^n \varphi$ on the one-dimensional space $\bigwedge^n V$ (Proposition 9.10.6 and Corollary 9.10.9). Every linear map L from a one-dimensional space to itself is multiplication by a scalar: if ω spans the space and $L(\omega) = c\,\omega$, then $L(x\omega) = c\,x\omega$. For $\bigwedge^n \varphi$, that scalar is the determinant.

Theorem 9.11.3 (The determinant via $\bigwedge^n V$)
Let V be a vector space of dimension $n \geqslant 1$ and $\varphi \in \mathrm{End}(V)$. Then $\bigwedge^n \varphi$ is multiplication by $\det(\varphi)$:

$$\varphi(v_1) \wedge \cdots \wedge \varphi(v_n) = \det(\varphi)\; v_1 \wedge \cdots \wedge v_n \qquad \text{for all } v_1, \dots, v_n \in V,$$

where $\det(\varphi)$ is the determinant of the matrix of φ in any basis (Definition 5.4.20). More generally, if $e_1, \dots, e_n$ is a basis of V and $v_j = \sum_i a_{ij} e_i$

for $j = 1, \dots, n$, then

$$v_1 \wedge \cdots \wedge v_n = \det(A)\, e_1 \wedge \cdots \wedge e_n, \qquad A = (a_{ij}).$$

Proof. Fix the basis $e_1, \dots, e_n$. Since $e_1 \wedge \cdots \wedge e_n$ is a basis of $\bigwedge^n V$, each $v_1 \wedge \cdots \wedge v_n$ is a unique multiple $D \cdot e_1 \wedge \cdots \wedge e_n$, and we regard D as a function of the $n \times n$ matrix A whose jth column is the coordinate vector of v_j. It is multilinear in the columns, because the wedge product is multilinear and taking the coefficient of $e_1 \wedge \cdots \wedge e_n$ is linear; it is alternating in the columns, because a wedge product with two equal factors vanishes; and $D(I) = 1$, since then $v_j = e_j$. By Theorem 5.4.4(b), $D(A) = D(I)\det(A) = \det(A)$. This proves the second statement.

Applying it to $v_j = \varphi(e_j)$, whose coordinate vectors are the columns of the matrix A of φ, gives $\bigwedge^n \varphi\,(e_1 \wedge \cdots \wedge e_n) = \det(A)\, e_1 \wedge \cdots \wedge e_n$. As $e_1 \wedge \cdots \wedge e_n$ spans $\bigwedge^n V$, $\bigwedge^n \varphi$ is multiplication by $\det(A)$, which is $\det(\varphi)$. □

NOTE The Leibniz formula can also be read off directly. With $v_j = \sum_i a_{ij} e_i$, multilinearity gives

$$v_1 \wedge \cdots \wedge v_n = \sum_{i_1, \dots, i_n} a_{i_1 1} a_{i_2 2} \cdots a_{i_n n}\; e_{i_1} \wedge \cdots \wedge e_{i_n}.$$

A term with a repeated index vanishes. The others have $(i_1, \dots, i_n) = (\sigma(1), \dots, \sigma(n))$ for a permutation $\sigma \in S_n$, and $e_{\sigma(1)} \wedge \cdots \wedge e_{\sigma(n)} = \operatorname{sgn}(\sigma)\, e_1 \wedge \cdots \wedge e_n$ by Lemma 9.10.2(ii). Hence

$$\det(A) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma)\, a_{\sigma(1)1}\, a_{\sigma(2)2} \cdots a_{\sigma(n)n},$$

which is the Leibniz formula of Definition 5.4.3.

NOTE We could have taken Theorem 9.11.3 as the definition: $\det(\varphi)$ is the scalar by which φ acts on the determinant line $\bigwedge^n V$. This definition needs no basis, no permutations and no cofactor expansions, so the independence of $\det(\varphi)$ from the basis is manifest. It also has a geometric meaning: $\bigwedge^n V$ is where signed volumes live, and $\det(\varphi)$ is the factor by which φ scales them (see the end of this section).

Corollary 9.11.4 (Properties of the determinant)
Let $\dim V = n \geqslant 1$, $\varphi, \psi \in \operatorname{End}(V)$ and $\lambda \in k$.

(i) $\det(\psi \circ \varphi) = \det(\psi)\det(\varphi)$.
(ii) $\det(\mathrm{id}_V) = 1$.
(iii) φ is invertible if and only if $\det(\varphi) \neq 0$.
(iv) $\det(\lambda\varphi) = \lambda^n \det(\varphi)$.

Proof. (i) $\bigwedge^n(\psi \circ \varphi) = \bigwedge^n \psi \circ \bigwedge^n \varphi$ (Proposition 9.10.6), and multiplication by $\det(\varphi)$ followed by multiplication by $\det(\psi)$ is multiplication by $\det(\psi)\det(\varphi)$.

(ii) $\bigwedge^n(\mathrm{id}_V)$ is the identity of $\bigwedge^n V$.

(iii) Let $e_1, \dots, e_n$ be a basis of V. Then φ is invertible if and only if $\varphi(e_1), \dots, \varphi(e_n)$ is a basis of V, that is, if and only if these n vectors are linearly independent; by Proposition 9.10.10, this happens if and only if $\varphi(e_1) \wedge \cdots \wedge \varphi(e_n) = \det(\varphi)\, e_1 \wedge \cdots \wedge e_n$ is nonzero.

(iv) $(\lambda\varphi)(v_1) \wedge \cdots \wedge (\lambda\varphi)(v_n) = \lambda^n\, \varphi(v_1) \wedge \cdots \wedge \varphi(v_n)$ by multilinearity. □

Applied to the operators of k^n given by two matrices A and B, part (i) gives $\det(AB) = \det(A)\det(B)$, which is Theorem 5.4.7, now without any computation; part (iii) is Theorem 5.4.8. Neither argument is circular: the proof of Theorem 9.11.3 used only the characterization of the determinant by its columns (Theorem 5.4.4), which does not depend on multiplicativity.

9.11.3 Exterior Powers of an Operator

The coefficients of a wedge product of fewer than n vectors are determinants as well.

Lemma 9.11.5 (Coordinates of a wedge product)

Let $e_1, \dots, e_n$ be a basis of V, let $1 \leqslant d \leqslant n$, and let $w_1, \dots, w_d \in V$ with $w_b = \sum_i a_{ib} e_i$, so that $A = (a_{ib})$ is an $n \times d$ matrix. For an increasing tuple $I = (i_1 < \cdots < i_d)$, let A_I be the $d \times d$ matrix formed by the rows $i_1, \dots, i_d$ of A. Then

$$w_1 \wedge \cdots \wedge w_d = \sum_I \det(A_I)\, e_I.$$

Proof. Let $p_I : V \to k^d$ send $\sum_i x_i e_i$ to $(x_{i_1}, \dots, x_{i_d})$, and let $\varepsilon_1, \dots, \varepsilon_d$ be the standard basis of k^d. The map $\bigwedge^d p_I$ sends e_J to $p_I(e_{j_1}) \wedge \cdots \wedge p_I(e_{j_d})$, which is $\varepsilon_1 \wedge \cdots \wedge \varepsilon_d$ if $J = I$ and 0 otherwise (if $J \neq I$, some j_b is not among the i_a, and $p_I(e_{j_b}) = 0$). So $\bigwedge^d p_I$ extracts the coefficient of e_I. On the other hand, the vectors $p_I(w_1), \dots, p_I(w_d)$ are the columns of A_I, so by Theorem 9.11.3 in k^d,

$$\textstyle\bigwedge^d p_I\,(w_1 \wedge \cdots \wedge w_d) = p_I(w_1) \wedge \cdots \wedge p_I(w_d) = \det(A_I)\ \varepsilon_1 \wedge \cdots \wedge \varepsilon_d. \quad \square$$

Combined with Proposition 9.10.10, the lemma says that the columns of an $n \times d$ matrix are linearly independent if and only if some $d \times d$ minor of the matrix is nonzero.

Proposition 9.11.6 (Traces of exterior powers and the characteristic polynomial)

Let $\dim V = n \geqslant 1$, and let $\varphi \in \mathrm{End}(V)$ have matrix A in a basis $e_1, \dots, e_n$. For an increasing d-tuple I, let A_{II} be the $d \times d$ submatrix of A with rows and columns $i_1, \dots, i_d$; its determinant is a ***principal minor*** of A.

(i) $\mathrm{Tr}(\bigwedge^d \varphi) = \sum_I \det(A_{II})$, the sum of the principal $d \times d$ minors of A.

(ii) The characteristic polynomial of φ (Definition 7.4.7) is

$$\chi_\varphi(x) = \det(xI - A) = \sum_{d=0}^{n} (-1)^d \ \mathrm{Tr}\left(\bigwedge^d \varphi\right) x^{n-d}.$$

(iii) If A is upper triangular with diagonal entries $\lambda_1, \dots, \lambda_n$, then $\mathrm{Tr}(\bigwedge^d \varphi) = \sum_{i_1 < \cdots < i_d} \lambda_{i_1} \cdots \lambda_{i_d}$, the dth elementary symmetric polynomial in $\lambda_1, \dots, \lambda_n$.

Proof. (i) The diagonal entry of the matrix of $\bigwedge^d \varphi$ in the basis (e_I) at position I is the coefficient of e_I in $\bigwedge^d \varphi\,(e_I) = \varphi(e_{i_1}) \wedge \cdots \wedge \varphi(e_{i_d})$. The coordinate vectors of $\varphi(e_{i_1}), \dots, \varphi(e_{i_d})$ are the columns $i_1, \dots, i_d$ of A, and rows $i_1, \dots, i_d$ of these columns form A_{II}; so the coefficient is $\det(A_{II})$ by Lemma 9.11.5. Summing over I gives the trace.

(ii) The determinant $\det(xI - A)$ is given by the Leibniz formula, so it is the same polynomial whether we regard $xI - A$ as a matrix with entries in $k[x]$ or in the field $K = k(x)$ of rational functions (Example 5.4.25). We work over K, in K^n with standard basis $\varepsilon_1, \dots, \varepsilon_n$. Let a_j be the jth column of A; the jth column of $xI - A$ is $x\varepsilon_j - a_j$. By Theorem 9.11.3 and multilinearity,

$$\begin{aligned} \det(xI - A)\ \varepsilon_1 \wedge \cdots \wedge \varepsilon_n &= (x\varepsilon_1 - a_1) \wedge \cdots \wedge (x\varepsilon_n - a_n) \\ &= \sum_{S \subseteq \{1,\dots,n\}} x^{|S|} (-1)^{n-|S|}\ w_1^S \wedge \cdots \wedge w_n^S, \end{aligned}$$

where $w_j^S = \varepsilon_j$ for $j \in S$ and $w_j^S = a_j$ for $j \notin S$. Fix S, and let I be the increasing tuple listing the complement of S, with $d = n - |S|$ entries. In $w_1^S \wedge \cdots \wedge w_n^S$ we may replace each a_j ($j \in I$) by $a_j' = \sum_{i \in I} a_{ij} \varepsilon_i$, because the rest of a_j is a combination of the ε_s with $s \in S$, which already occur as factors. Let $\sigma \in S_n$ list the elements $s_1 < \cdots < s_{n-d}$ of S followed by $i_1, \dots, i_d$. Reordering the factors according to σ multiplies both $w_1^S \wedge \cdots \wedge w_n^S$ and $\varepsilon_1 \wedge \cdots \wedge \varepsilon_n$ by $\mathrm{sgn}(\sigma)$

(Lemma 9.10.2(ii)). After reordering, the wedge product is

$$\varepsilon_{s_1} \wedge \cdots \wedge \varepsilon_{s_{n-d}} \wedge a'_{i_1} \wedge \cdots \wedge a'_{i_d} = \det(A_{II})\ \varepsilon_{s_1} \wedge \cdots \wedge \varepsilon_{s_{n-d}} \wedge \varepsilon_{i_1} \wedge \cdots \wedge \varepsilon_{i_d},$$

by Lemma 9.11.5: the coordinates of $a'_{i_1}, \dots, a'_{i_d}$ vanish outside the rows in I, and in those rows they form A_{II}. Hence $w_1^S \wedge \cdots \wedge w_n^S = \det(A_{II})\, \varepsilon_1 \wedge \cdots \wedge \varepsilon_n$, and

$$\det(xI - A) = \sum_{d=0}^{n} (-1)^d\, x^{n-d} \sum_{|I|=d} \det(A_{II}) = \sum_{d=0}^{n} (-1)^d\ \operatorname{Tr}\left(\bigwedge^d \varphi\right) x^{n-d}$$

by (i). (For $d = 0$ the inner sum is 1, and $\bigwedge^0 \varphi = \mathrm{id}_k$ has trace 1.)

(iii) Each A_{II} is upper triangular with diagonal entries $\lambda_{i_1}, \dots, \lambda_{i_d}$, so its determinant is their product (Proposition 5.4.10); apply (i). □

When k is algebraically closed, Theorem 7.5.2 provides a basis in which A is upper triangular; then $xI - A$ is upper triangular too, and $\chi_\varphi(x) = \det(xI - A) = \prod_i (x - \lambda_i)$, so the λ_i are the eigenvalues of φ, each repeated according to its algebraic multiplicity, and (iii) says that $\operatorname{Tr}(\bigwedge^d \varphi)$ is the dth elementary symmetric polynomial in the eigenvalues. The cases $d = 1$ and $d = n$ are $\operatorname{Tr}(\varphi) = \sum_i \lambda_i$ and $\det(\varphi) = \prod_i \lambda_i$.

9.11.4 Volume Forms and Orientation

Let $\dim V = n \geqslant 1$. Over $\mathbb{R}$, think of the wedge product $v_1 \wedge \cdots \wedge v_n$ as the signed volume of the parallelepiped $\{t_1 v_1 + \cdots + t_n v_n : 0 \leqslant t_i \leqslant 1\}$. It behaves as a signed volume should: it is multilinear and alternating, and it vanishes exactly when the parallelepiped is flat, that is, when the v_i are dependent (Proposition 9.10.10). But it lives in the line $\bigwedge^n V$, and turning it into a number requires an isomorphism $\bigwedge^n V \to k$.

Definition 9.11.7 (Volume form)

Let $\dim V = n \geqslant 1$. A ***volume form*** on V is a nonzero alternating n-linear form $\mathrm{vol} : V^n \to k$. By Proposition 9.10.5, volume forms correspond to the nonzero elements L of $(\bigwedge^n V)^*$, that is, to the isomorphisms $L : \bigwedge^n V \to k$, via $\mathrm{vol}(v_1, \dots, v_n) = L(v_1 \wedge \cdots \wedge v_n)$.

Since $(\bigwedge^n V)^*$ is one-dimensional, any two volume forms differ by a nonzero scalar factor. A vector space has no preferred volume form until we give it more structure, such as an inner product.

Proposition 9.11.8 (Volume forms and determinants)

Let vol be a volume form on V.

(i) $\mathrm{vol}(v_1,\dots,v_n) \neq 0$ if and only if $v_1,\dots,v_n$ is a basis of V.

(ii) $\mathrm{vol}(\varphi(v_1),\dots,\varphi(v_n)) = \det(\varphi)\,\mathrm{vol}(v_1,\dots,v_n)$ for every $\varphi \in \mathrm{End}(V)$.

(iii) If $e_1,\dots,e_n$ is a basis with $\mathrm{vol}(e_1,\dots,e_n) = 1$, then $\mathrm{vol}(v_1,\dots,v_n) = \det(A)$, where the jth column of A is the coordinate vector of v_j in this basis.

Proof. Write $\mathrm{vol}(v_1,\dots,v_n) = L(v_1 \wedge \cdots \wedge v_n)$ with L an isomorphism. (i) follows from Proposition 9.10.10, since n independent vectors in V form a basis. (ii) and (iii) follow by applying L to the two formulas of Theorem 9.11.3; in (iii), $L(e_1 \wedge \cdots \wedge e_n) = 1$. □

So $\det(\varphi)$ is the factor by which φ multiplies all signed volumes, whichever volume form is used to measure them. Volume forms can also be written as wedge products of linear functionals.

Proposition 9.11.9 (Exterior powers of the dual space)

For $1 \leqslant d \leqslant n$ there is an isomorphism $\bigwedge^d(V^*) \to (\bigwedge^d V)^*$ under which $\ell_1 \wedge \cdots \wedge \ell_d$ corresponds to the functional

$$v_1 \wedge \cdots \wedge v_d \mapsto \det\left(\ell_i(v_j)\right)_{i,j=1}^{d}.$$

If $e_1,\dots,e_n$ is a basis of V, it sends $e_I^* = e_{i_1}^* \wedge \cdots \wedge e_{i_d}^*$ to the dual basis of the basis (e_I) of $\bigwedge^d V$. In particular, the alternating d-linear forms on V are the elements of $\bigwedge^d(V^*)$.

Proof. For fixed $\ell_1,\dots,\ell_d$, the function $(v_1,\dots,v_d) \mapsto \det(\ell_i(v_j))$ is multilinear and alternating, because v_j enters only the jth column of the matrix and the determinant is multilinear and alternating in the columns (Theorem 5.4.4(a)); so it defines an element f_ℓ of $(\bigwedge^d V)^*$. For fixed $v_1,\dots,v_d$, the value is multilinear and alternating in the ℓ_i, because ℓ_i enters only the ith row and the determinant is also multilinear and alternating in the rows (Proposition 5.4.6). Since f_ℓ is determined by its values on the spanning set of wedge products, $(\ell_1,\dots,\ell_d) \mapsto f_\ell$ is multilinear and alternating, and it induces a linear map $\bigwedge^d(V^*) \to (\bigwedge^d V)^*$. For increasing tuples I and J, the value of $f_{e_I^*}$ on e_J is $\det(e_{i_a}^*(e_{j_b}))_{a,b}$. This is $\det(I_d) = 1$ when $I = J$; when $I \neq J$, some i_a is not among the j_b, the ath row is zero, and the determinant is 0. So the basis (e_I^*) of $\bigwedge^d(V^*)$ is sent to the dual basis of (e_I), and the map is an isomorphism. The last statement follows from Proposition 9.10.5. □

Example 9.11.10 (The volume form of a real inner product space)

Let V be a real inner product space (Definition 9.1.11) with an orthonormal basis $e_1, \dots, e_n$. The volume form $\mathrm{vol} = e_1^* \wedge \cdots \wedge e_n^*$ satisfies $\mathrm{vol}(e_1, \dots, e_n) = \det(I) = 1$: the unit cube has volume 1. By Proposition 9.11.8(iii), $\mathrm{vol}(v_1, \dots, v_n)$ is the determinant of the matrix of coordinates of the v_j.

Let $e'_1, \dots, e'_n$ be another orthonormal basis, with $e'_j = \sum_i p_{ij} e_i$. Then $(P^t P)_{ij} = \sum_k p_{ki} p_{kj} = \langle e'_i, e'_j \rangle = \delta_{ij}$, so P is orthogonal, and $\det(P)^2 = \det(P^t)\det(P) = 1$ by Proposition 5.4.6 and Theorem 5.4.7. Hence $\mathrm{vol}(e'_1, \dots, e'_n) = \det(P) = \pm 1$, and the volume form built from (e'_i) is $\pm\mathrm{vol}$. So the inner product determines the volume form up to sign. Reordering an orthonormal basis can already change the sign: the basis $e_2, e_1, e_3, \dots, e_n$ gives $-\mathrm{vol}$. Choosing the sign is choosing an orientation.

Definition 9.11.11 (Orientation)

Let V be a real vector space of dimension $n \geqslant 1$. Two ordered bases $(e_1, \dots, e_n)$ and $(e'_1, \dots, e'_n)$ of V have the ***same orientation*** if $e'_1 \wedge \cdots \wedge e'_n = c\, e_1 \wedge \cdots \wedge e_n$ with $c > 0$. An ***orientation*** of V is an equivalence class of ordered bases under this relation, and the bases in the chosen class are called ***positively oriented***.

Proposition 9.11.12 (Orientations)

Let V be a real vector space of dimension $n \geqslant 1$.

(i) Having the same orientation is an equivalence relation with exactly two classes. The constant c of Definition 9.11.11 is $\det(P)$, where $e'_j = \sum_i p_{ij} e_i$.

(ii) An orientation amounts to a choice of one of the two half-lines of $\bigwedge^n V \setminus \{0\}$, or of a volume form up to multiplication by a positive scalar.

(iii) An invertible $\varphi \in \mathrm{End}(V)$ sends positively oriented bases to positively oriented bases if and only if $\det(\varphi) > 0$.

Proof. (i) The formula $c = \det(P)$ is Theorem 9.11.3. The relation is reflexive ($c = 1$), symmetric (c becomes $1/c$) and transitive (the constants multiply), since inverses and products of positive numbers are positive. Fix a basis (e_i). For every basis (e'_i), the constant c is nonzero (Proposition 9.10.10), so it is positive or negative, and both signs occur, because replacing e'_1 by $-e'_1$ changes the sign of c. So there are exactly two classes.

(ii) The class of (e_i') is determined by the half-line of $\bigwedge^n V \setminus \{0\}$ containing $e_1' \wedge \cdots \wedge e_n'$. A volume form vol determines the class of the bases with $\mathrm{vol}(e_1', \ldots, e_n') > 0$, and vol and $c\,\mathrm{vol}$ determine the same class if and only if $c > 0$.

(iii) $\varphi(e_1') \wedge \cdots \wedge \varphi(e_n') = \det(\varphi)\, e_1' \wedge \cdots \wedge e_n'$. □

Example 9.11.13 (Orientations of the plane)

In $\mathbb{R}^2$, the bases (e_1, e_2) and (e_2, e_1) have opposite orientations, since $e_2 \wedge e_1 = -e_1 \wedge e_2$. Rotations have determinant 1 and preserve orientation; reflections have determinant -1 and reverse it, which matches the description of $\mathrm{O}(2)$ in Proposition 9.5.16.

Exercise 9.11.14 (Tensor products and multilinear algebra exercises)

(1) Let V be a two-dimensional vector space over k with basis e_1, e_2.

   (a) Write down a basis of $V \otimes V$, and find bases of the subspaces $\mathcal{S}^2(V)$ and $\mathcal{A}^2(V)$ of symmetric and alternating tensors when $\mathrm{char}\, k \neq 2$. Verify that $\dim \mathcal{S}^2(V) + \dim \mathcal{A}^2(V) = \dim(V \otimes V)$, and find the alternating tensor corresponding to $e_1 \wedge e_2$ under Proposition 9.10.12.

   (b) Suppose $\mathrm{char}\, k = 2$. Show that $\mathcal{A}^2(V) = \mathcal{S}^2(V)$ is three-dimensional, whereas $\dim \bigwedge^2 V = 1$.

(2) Starting from the expression for $\det(A)$ obtained from wedge products in the note following Theorem 9.11.3, prove that $\det(A^t) = \det(A)$ for every $n \times n$ matrix A.

(3) Let $\dim V = n \geqslant 1$.

   (a) Show that $\eta \mapsto (v \mapsto v \wedge \eta)$ is an isomorphism $\bigwedge^{n-1} V \to \mathrm{Hom}(V, \bigwedge^n V) \cong V^* \otimes \bigwedge^n V$; it involves no choices.

   (b) Deduce that each nonzero $\omega \in \bigwedge^n V$ determines an isomorphism $\Phi_\omega : \bigwedge^{n-1} V \to V^*$, where $\Phi_\omega(\eta)(v) = c$ when $v \wedge \eta = c\,\omega$. How does Φ_ω change when ω is replaced by $\lambda\omega$?

(4) Let $V = \mathbb{R}^3$ with the standard inner product and standard basis e_1, e_2, e_3, and let $\omega = e_1 \wedge e_2 \wedge e_3$. Compose Φ_ω from (3) with the isomorphism $V^* \to V$ given by the inner product ($\ell \mapsto u$, where $\ell = \langle u, \cdot \rangle$). Show that the resulting isomorphism $\bigwedge^2 V \to V$ sends $v \wedge w$ to the cross product $v \times w$.

Solution. **(1)** (a) A basis of $V \otimes V$ is $e_1 \otimes e_1,\ e_1 \otimes e_2,\ e_2 \otimes e_1,\ e_2 \otimes e_2$, so $\dim(V \otimes V) = 4$. The group S_2 has one nontrivial element, acting by the swap τ. For $\eta = \sum_{i,j} c_{ij}\, e_i \otimes e_j$ we have $\tau\eta = \sum_{i,j} c_{ij}\, e_j \otimes e_i$, so η is symmetric if and only if $c_{12} = c_{21}$. A basis of $\mathcal{S}^2(V)$ is

$$e_1 \otimes e_1, \qquad e_2 \otimes e_2, \qquad e_1 \otimes e_2 + e_2 \otimes e_1,$$

and $\dim \mathcal{S}^2(V) = 3 = \binom{2+2-1}{2}$. The tensor η is alternating if and only if $\tau\eta = -\eta$, that is, $c_{11} = -c_{11}$, $c_{22} = -c_{22}$ and $c_{21} = -c_{12}$. When $\operatorname{char} k \neq 2$ this means $c_{11} = c_{22} = 0$ and $c_{21} = -c_{12}$, so $\mathcal{A}^2(V)$ has basis $e_1 \otimes e_2 - e_2 \otimes e_1$ and dimension $1 = \binom{2}{2}$. Indeed $3+1 = 4$, as Corollary 9.10.13 predicts. Under Proposition 9.10.12, $e_1 \wedge e_2$ corresponds to $P_{\mathcal{A}}(e_1 \otimes e_2) = \frac{1}{2}(e_1 \otimes e_2 - e_2 \otimes e_1)$.
(b) In characteristic 2 we have $-1 = 1$, so the condition $\tau\eta = -\eta$ is the condition $\tau\eta = \eta$: $\mathcal{A}^2(V) = \mathcal{S}^2(V)$, which is three-dimensional by the computation in (a) (that computation did not use the characteristic). But $\dim \bigwedge^2 V = \binom{2}{2} = 1$ by Proposition 9.10.8, which holds in every characteristic. So the subspace of alternating tensors is not a model of $\bigwedge^2 V$ in characteristic 2. The subspace of symmetric tensors fails as a model of $\operatorname{Sym}^2 V$ as well, although its dimension is right: the quotient map sends $e_1 \otimes e_2 + e_2 \otimes e_1$ to $2\, e_1 e_2 = 0$, so it is not injective on $\mathcal{S}^2(V)$.
(2) By the note following Theorem 9.11.3,

$$\det(A) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{j=1}^{n} a_{\sigma(j)j}.$$

The (i, j) entry of A^t is a_{ji}, so

$$\det(A^t) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{j=1}^{n} a_{j\sigma(j)}.$$

For fixed σ, put $\tau = \sigma^{-1}$. As j runs over $1, \dots, n$, so does $i = \sigma(j)$, and $j = \tau(i)$; hence $\prod_j a_{j\sigma(j)} = \prod_i a_{\tau(i)i}$. Since $\operatorname{sgn}(\tau) = \operatorname{sgn}(\sigma)$ and $\sigma \mapsto \sigma^{-1}$ is a bijection of S_n,

$$\det(A^t) = \sum_{\tau \in S_n} \operatorname{sgn}(\tau) \prod_{i=1}^{n} a_{\tau(i)i} = \det(A).$$

(3) (a) The product $(\eta, \nu) \mapsto \nu \wedge \eta$ from $\bigwedge^{n-1} V \times V$ to $\bigwedge^n V$ is bilinear (Definition 9.10.17). So for each η the map $\Psi(\eta) : \nu \mapsto \nu \wedge \eta$ is linear, and $\eta \mapsto \Psi(\eta)$ is a linear map $\Psi : \bigwedge^{n-1} V \to \operatorname{Hom}(V, \bigwedge^n V)$. No choices were made. To see that Ψ is injective, choose a basis $e_1, \dots, e_n$. The increasing

$(n-1)$-tuples are the tuples $I_j = (1,\dots,\widehat{j},\dots,n)$ that omit one index j. Write $\eta = \sum_j c_j\, e_{I_j}$. For $j' \neq j$, $e_j \wedge e_{I_{j'}} = 0$, because j occurs in $I_{j'}$; and $e_j \wedge e_{I_j} = (-1)^{j-1} e_1 \wedge \cdots \wedge e_n$, by moving e_j past $j-1$ factors. So $\Psi(\eta)(e_j) = (-1)^{j-1} c_j\, e_1 \wedge \cdots \wedge e_n$, and $\Psi(\eta) = 0$ forces every $c_j = 0$. Both spaces have dimension n (Proposition 9.10.8), so Ψ is an isomorphism. Finally $\mathrm{Hom}(V, \bigwedge^n V) \cong V^* \otimes \bigwedge^n V$ by Proposition 9.8.10.
(b) A nonzero ω gives an isomorphism $\bigwedge^n V \to k$, $c\,\omega \mapsto c$, and hence $\mathrm{Hom}(V, \bigwedge^n V) \cong \mathrm{Hom}(V,k) = V^*$. Composing with (a) gives Φ_ω. Since $v \wedge \eta = c\,\omega = (c/\lambda)(\lambda\omega)$, we get $\Phi_{\lambda\omega} = \lambda^{-1}\Phi_\omega$. So the isomorphism $\bigwedge^{n-1} V \cong V^*$ depends on the choice of ω; the isomorphism of (a) does not.
(4) Let $v = \sum_i v_i e_i$ and $w = \sum_i w_i e_i$. Using $e_i \wedge e_i = 0$ and $e_j \wedge e_i = -e_i \wedge e_j$,

$$v \wedge w = (v_1 w_2 - v_2 w_1)\, e_1 \wedge e_2 + (v_1 w_3 - v_3 w_1)\, e_1 \wedge e_3 + (v_2 w_3 - v_3 w_2)\, e_2 \wedge e_3.$$

For $u = \sum_i u_i e_i$ we have $u \wedge e_2 \wedge e_3 = u_1\,\omega$, so $\Phi_\omega(e_2 \wedge e_3) = e_1^*$, which corresponds to e_1 because $\langle e_1, u\rangle = u_1$. Similarly $u \wedge e_1 \wedge e_3 = u_2\, e_2 \wedge e_1 \wedge e_3 = -u_2\,\omega$, so $e_1 \wedge e_3 \mapsto -e_2$; and $u \wedge e_1 \wedge e_2 = u_3\, e_3 \wedge e_1 \wedge e_2 = u_3\,\omega$ (moving e_3 past two factors), so $e_1 \wedge e_2 \mapsto e_3$. Therefore

$$v \wedge w \mapsto (v_2 w_3 - v_3 w_2)\, e_1 - (v_1 w_3 - v_3 w_1)\, e_2 + (v_1 w_2 - v_2 w_1)\, e_3 = v \times w.$$

The construction needs $\dim \bigwedge^2 V = \dim V$, that is, $\binom{n}{2} = n$, which for $n \geqslant 1$ holds only when $n = 3$. This is why it produces a cross product only in dimension 3. (A bilinear cross product also exists on $\mathbb{R}^7$, but it comes from a different construction.)

9.12 Interlude: Tensors in Differential Geometry and Physics

The tensors we have constructed belong to multilinear algebra: they live on a single fixed vector space. In differential geometry and physics one works with tensor *fields*, tensors that vary smoothly from point to point of a manifold. This section is a brief account of how the algebra above connects to the classical tensor calculus of Ricci, Levi-Civita, and Einstein. We do not develop manifolds in this book, so the geometric parts are descriptive.

9.12.1 Tensors of Type (r, s)

Let V be an n-dimensional real vector space with basis $e_1, \dots, e_n$. In this section we write the dual basis with upper indices, $e^1, \dots, e^n$, so that $e^i(e_j) = \delta^i_j$.

Definition 9.12.1 (Tensor of type (r, s))

A ***tensor of type*** (r, s) on V is an element of

$$T^r_s(V) = \underbrace{V \otimes \cdots \otimes V}_{r} \otimes \underbrace{V^* \otimes \cdots \otimes V^*}_{s}.$$

Using the isomorphism $V \cong V^{**}$ of Theorem 6.2.19 on the first r factors and then Proposition 9.9.6,

$$T^r_s(V) \cong (V^{**})^{\otimes r} \otimes (V^*)^{\otimes s} \cong \left((V^*)^{\otimes r} \otimes V^{\otimes s}\right)^*,$$

so a tensor of type (r, s) is the same as an $(r+s)$-linear form

$$T : \underbrace{V^* \times \cdots \times V^*}_{r} \times \underbrace{V \times \cdots \times V}_{s} \to \mathbb{R}.$$

A pure tensor $v_1 \otimes \cdots \otimes v_r \otimes \ell_1 \otimes \cdots \otimes \ell_s$ corresponds to the form

$$(\alpha_1, \ldots, \alpha_r, w_1, \ldots, w_s) \mapsto \alpha_1(v_1) \cdots \alpha_r(v_r)\, \ell_1(w_1) \cdots \ell_s(w_s).$$

In coordinates, T is specified by its n^{r+s} ***components***

$$T^{i_1 \cdots i_r}{}_{j_1 \cdots j_s} = T(e^{i_1}, \ldots, e^{i_r}, e_{j_1}, \ldots, e_{j_s}),$$

and it is recovered from them as

$$T = \sum T^{i_1 \cdots i_r}{}_{j_1 \cdots j_s}\; e_{i_1} \otimes \cdots \otimes e_{i_r} \otimes e^{j_1} \otimes \cdots \otimes e^{j_s},$$

because these tensors form a basis of $T^r_s(V)$ (Theorem 9.9.2), and the form corresponding to each of them takes the value 1 on one tuple of basis vectors and 0 on all the others. We use the ***Einstein summation convention***: an index that appears once as a superscript and once as a subscript is summed over. With it, the formula above reads $T = T^{i_1 \cdots i_r}{}_{j_1 \cdots j_s}\, e_{i_1} \otimes \cdots \otimes e_{i_r} \otimes e^{j_1} \otimes \cdots \otimes e^{j_s}$.

Example 9.12.2 (Familiar objects as tensors)

- A vector $v \in V$ is a tensor of type $(1, 0)$: $v = v^i e_i$.
- A covector (linear functional) $\omega \in V^*$ has type $(0, 1)$: $\omega = \omega_i\, e^i$.
- A bilinear form $b \in V^* \otimes V^*$ has type $(0, 2)$: $b = b_{ij}\, e^i \otimes e^j$, with $b_{ij} = b(e_i, e_j)$.

- A linear map $T : V \to V$ has type $(1,1)$, via $\mathrm{End}(V) \cong V^* \otimes V \cong V \otimes V^*$ (Propositions 9.8.10 and 9.8.8): $T = T^i{}_j \, e_i \otimes e^j$, where $T(e_j) = T^i{}_j \, e_i$. So $T^i{}_j$ is the (i,j) entry of the matrix of T.

9.12.2 Change of Basis and the Transformation Law

The defining feature of tensors in physics, and the historical reason physicists care about them, is the way their components change under a change of basis. Let $e'_1, \dots, e'_n$ be a new basis with $e'_i = A^j{}_i \, e_j$, so that $A = (A^j{}_i)$ is the change-of-basis matrix (row index j, column index i). The new dual basis is $e'^i = (A^{-1})^i{}_j \, e^j$, since $(A^{-1})^i{}_j \, e^j(A^l{}_k \, e_l) = (A^{-1})^i{}_j A^j{}_k = \delta^i_k$.

Proposition 9.12.3 (Tensor transformation law)

Under the change of basis $e'_i = A^j{}_i \, e_j$, the components of a tensor of type (r,s) transform as

$$T'^{i_1 \cdots i_r}{}_{j_1 \cdots j_s} = (A^{-1})^{i_1}{}_{k_1} \cdots (A^{-1})^{i_r}{}_{k_r} \, A^{l_1}{}_{j_1} \cdots A^{l_s}{}_{j_s} \, T^{k_1 \cdots k_r}{}_{l_1 \cdots l_s}.$$

Proof. By definition $T'^{i_1 \cdots i_r}{}_{j_1 \cdots j_s} = T(e'^{i_1}, \dots, e'^{i_r}, e'_{j_1}, \dots, e'_{j_s})$. Substitute $e'^i = (A^{-1})^i{}_k \, e^k$ and $e'_j = A^l{}_j \, e_l$, and expand by multilinearity. □

NOTE The transformation law expresses the fact that the tensor T itself, as an element of $T^r_s(V)$, does not depend on a basis; only its components do. Many older texts *defined* a tensor as a collection of numbers $T^{i_1 \cdots i_r}{}_{j_1 \cdots j_s}$, one for each basis, related by this law. With tensor products, the law becomes a theorem.

9.12.3 Tensor Fields on Manifolds

A smooth manifold M is a space that looks locally like $\mathbb{R}^n$, such as a sphere or a torus. Each point $p \in M$ has a ***tangent space*** T_pM, a real vector space of dimension $\dim M$. A ***tensor field of type*** (r,s) on M is a smooth assignment $p \mapsto T_p$, where $T_p \in T^r_s(T_pM)$ for each p.

Local coordinates $x^1, \dots, x^n$ near p give a basis of the tangent space T_pM, written $\partial/\partial x^1, \dots, \partial/\partial x^n$, and the dual basis $dx^1, \dots, dx^n$ of its dual space, the cotangent space T^*_pM. A tensor field is then

$$T = T^{i_1 \cdots i_r}{}_{j_1 \cdots j_s}(x) \, \frac{\partial}{\partial x^{i_1}} \otimes \cdots \otimes \frac{\partial}{\partial x^{i_r}} \otimes dx^{j_1} \otimes \cdots \otimes dx^{j_s},$$

where the component functions $T^{i_1\cdots i_r}{}_{j_1\cdots j_s}(x)$ are smooth. If $x'^1,\dots,x'^n$ are other local coordinates, the chain rule gives

$$\frac{\partial}{\partial x'^i} = \frac{\partial x^j}{\partial x'^i}\,\frac{\partial}{\partial x^j}, \qquad dx'^i = \frac{\partial x'^i}{\partial x^j}\,dx^j.$$

So the components transform by Proposition 9.12.3 with $A^j{}_i = \partial x^j/\partial x'^i$ and $(A^{-1})^i{}_j = \partial x'^i/\partial x^j$. For a vector field this gives the familiar rule $v'^i = (\partial x'^i/\partial x^j)\,v^j$, and for a covector field $\omega'_i = (\partial x^j/\partial x'^i)\,\omega_j$.

9.12.4 Index Gymnastics: Raising, Lowering, and Contraction

Suppose V carries a nondegenerate symmetric bilinear form g, such as an inner product or the metric tensor below. Write $g_{ij} = g(e_i, e_j)$, and let g^{ij} be the entries of the inverse matrix, so that $g^{ik}g_{kj} = \delta^i_j$. Then indices can be "raised" and "lowered".

Definition 9.12.4 (Raising and lowering indices)

Given a tensor $T^i{}_j$ of type $(1,1)$:

- ***Lowering*** the upper index gives $T_{ij} = g_{ik}\,T^k{}_j$, of type $(0,2)$.
- ***Raising*** the lower index gives $T^{ij} = g^{jk}\,T^i{}_k$, of type $(2,0)$.

In general, any index can be raised with g^{ij} or lowered with g_{ij}.

In coordinate-free terms, lowering an index applies the map $\varphi_g : V \to V^*$, $v \mapsto g(v,\cdot\,)$, of Theorem 9.1.8 (an isomorphism because g is nondegenerate, Definition 9.1.9) to one tensor factor, and raising applies its inverse. For a vector, lowering gives $v_i = g_{ij}v^j$.

Definition 9.12.5 (Contraction of a tensor)

Let $r, s \geqslant 1$, $1 \leqslant p \leqslant r$ and $1 \leqslant q \leqslant s$. The ***contraction*** of the pth upper index with the qth lower index is the linear map $C^p_q : T^r_s(V) \to T^{r-1}_{s-1}(V)$ with

$$C^p_q(v_1\otimes\cdots\otimes v_r\otimes \ell_1\otimes\cdots\otimes\ell_s) = \ell_q(v_p)\;v_1\otimes\cdots\widehat{v_p}\cdots\otimes v_r\otimes\ell_1\otimes\cdots\widehat{\ell_q}\cdots\otimes\ell_s,$$

where a hat marks an omitted factor. In components,

$$(C^p_q T)^{i_1\cdots\widehat{i_p}\cdots i_r}{}_{j_1\cdots\widehat{j_q}\cdots j_s} = T^{i_1\cdots k\cdots i_r}{}_{j_1\cdots k\cdots j_s},$$

with k in the pth upper and the qth lower position, summed over.

The map C^p_q exists by the universal property, because the right side of its defining formula is multilinear in the v_i and ℓ_j; so contraction does not depend on a basis. The component formula follows by applying C^p_q to the basis tensors and using $e^j(e_i) = \delta^j_i$. For type $(1,1)$, contraction is the map c of Definition 9.11.1 (with the two factors swapped), and contracting $T^i{}_j$ gives $T^i{}_i = \operatorname{Tr}(T)$ by Proposition 9.11.2.

9.12.5 The Metric Tensor and General Relativity

The central object of Riemannian geometry is the ***metric tensor***: a smooth tensor field g of type $(0,2)$ whose value g_p at each point is an inner product on T_pM. In local coordinates,

$$g = g_{ij}(x)\, dx^i \otimes dx^j,$$

where $(g_{ij}(x))$ is a symmetric positive-definite matrix at each point x. The metric determines lengths, angles, volumes, curvature, and geodesics.

In Einstein's general relativity, spacetime is a four-dimensional ***pseudo-Riemannian manifold*** (M, g): each g_p is only required to be a nondegenerate symmetric bilinear form, and in general relativity it has signature $(1,3)$ (Definition 9.6.10). That is, in a basis that diagonalizes g_p, one diagonal entry is positive and three are negative, and by Sylvester's law of inertia (Theorem 9.6.9) these numbers do not depend on the basis. (Some authors use the opposite sign convention, signature $(3,1)$.) So each tangent space, with g_p, looks like Minkowski space: $\mathbb{R}^4$ with the form $x_0y_0 - x_1y_1 - x_2y_2 - x_3y_3$. The ***Einstein field equations***

$$G_{\mu\nu} + \Lambda\, g_{\mu\nu} = \frac{8\pi G}{c^4}\, T_{\mu\nu}$$

relate the geometry of spacetime, encoded in the ***Einstein tensor*** $G_{\mu\nu}$ (a tensor field of type $(0,2)$ built from g and its first and second derivatives), to its matter and energy content, encoded in the ***stress–energy tensor*** $T_{\mu\nu}$, also of type $(0,2)$. Here Λ is the cosmological constant (it has nothing to do with exterior powers), G is Newton's gravitational constant, and c is the speed of light.

> NOTE The Einstein tensor is $G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}R\, g_{\mu\nu}$, where $R_{\mu\nu}$ is the Ricci curvature tensor, a contraction of the Riemann curvature tensor $R^\rho{}_{\sigma\mu\nu}$ of type $(1,3)$, and $R = g^{\mu\nu}R_{\mu\nu}$ is the scalar curvature, a further contraction. Tensor products, contraction, and raising and lowering indices with the metric are exactly the operations needed to write these equations.

Example 9.12.6 (Contraction in general relativity)

The Riemann curvature tensor $R^{\rho}{}_{\sigma\mu\nu}$ has type $(1,3)$. Contracting its upper index with its second lower index gives the Ricci tensor $R_{\sigma\nu} = R^{\mu}{}_{\sigma\mu\nu}$, of type $(0,2)$. Raising an index with the inverse metric and contracting again gives the scalar curvature $R = g^{\sigma\nu}R_{\sigma\nu}$. Each contraction removes one upper and one lower index. The Riemann tensor carries all the curvature information; the Ricci tensor measures how the volumes of small regions are distorted compared with flat space; and the scalar curvature is a single number at each point, an average of the curvature there.

NOTE The whole apparatus of tensor calculus (index notation, Einstein summation, raising and lowering, contraction) is the coordinate form of the multilinear algebra developed in this chapter. The coordinate-free approach gives conceptual clarity and makes independence of the basis automatic; index notation is efficient for computation. Both perspectives are essential.

PART III

The Structure of Groups

CHAPTER 10

Modules and Finitely Generated Abelian Groups

A vector space is an abelian group on which the elements of a field act by scalar multiplication. If we let the scalars come from a ring instead, we obtain a *module*. Some of linear algebra survives this change and much of it does not; the first section of this chapter explores both. The second section specializes to the ring $\mathbb{Z}$. Every abelian group $(G, +)$ carries a multiplication by integers: for $n > 0$ let $ng = g + g + \cdots + g$ (n terms), let $0g = 0$, and let $ng = -(|n|g)$ for $n < 0$. This makes G a module over $\mathbb{Z}$, and the classification of finitely generated abelian groups becomes a problem of linear algebra over $\mathbb{Z}$.

10.1 Modules over a Ring

10.1.1 Definitions and Examples

Definition 10.1.1 (Module)

Let R be a ring (Definition 4.1.1). An R-***module*** is an abelian group $(M, +)$ together with a scalar multiplication $R \times M \to M$, $(r, m) \mapsto rm$, such that for all $r, r_1, r_2 \in R$ and $m, m_1, m_2 \in M$:

(i) $r(m_1 + m_2) = rm_1 + rm_2$;
(ii) $(r_1 + r_2)m = r_1m + r_2m$;
(iii) $(r_1r_2)m = r_1(r_2m)$;
(iv) $1m = m$.

As for vector spaces, $0m = 0$ and $(-1)m = -m$ follow from the axioms: $0m = (0+0)m = 0m+0m$ gives $0m = 0$, and then $m+(-1)m = (1+(-1))m = 0$. When R is not commutative, a module in this sense is also called a *left* R-module, since the scalars act on the left. From now on, R denotes a commutative ring.

Example 10.1.2 (Vector spaces)

If $R = k$ is a field, a k-module is exactly a vector space over k (Definition 4.5.1). Modules generalize vector spaces by allowing the scalars to come from a ring.

Example 10.1.3 (The modules R^n)

The set $R^n = \{(x_1, \ldots, x_n) : x_i \in R\}$, with componentwise addition and $r(x_1, \ldots, x_n) = (rx_1, \ldots, rx_n)$, is an R-module. For $n = 1$ this says that R is a module over itself, the scalar multiplication being the multiplication of R.

Example 10.1.4 (Abelian groups as $\mathbb{Z}$-modules)

Every abelian group $(G, +)$ is a $\mathbb{Z}$-module, with ng defined as in the introduction. The axioms are the rules $n(g + h) = ng + nh$, $(m + n)g = mg + ng$, and $(mn)g = m(ng)$, which follow from the definition by induction on $|m|$ and $|n|$ (the first uses that G is abelian). Conversely, a $\mathbb{Z}$-module structure on an abelian group is forced: by axioms (ii) and (iv), $ng = (1 + \cdots + 1)g = g + \cdots + g$ for $n > 0$, and $(-n)g = -(ng)$. So a $\mathbb{Z}$-module is the same thing as an abelian group.

Example 10.1.5 (Restriction of scalars)

If $\varphi : R \to S$ is a ring homomorphism, then any S-module M becomes an R-module via $rm = \varphi(r)m$. This is called ***restriction of scalars***. For instance, every $k[x]$-module is a vector space over k, by restriction along the inclusion $k \hookrightarrow k[x]$.

Example 10.1.6 (Modules over $k[x]$)

Let V be a vector space over k and $T : V \to V$ a linear operator. Then V becomes a $k[x]$-module via $p(x) \cdot v = p(T)(v)$; the axioms hold because $(p + q)(T) = p(T) + q(T)$ and $(pq)(T) = p(T) \circ q(T)$. Conversely, a $k[x]$-module V is a vector space over k (by restriction of scalars) together with the operator $T(v) = x \cdot v$. So a $k[x]$-module is the same thing as a vector space with an operator, and different operators give different module structures on the same vector space.

10.1.2 Submodules, Ideals, and Quotients

Definition 10.1.7 (Submodule)

A ***submodule*** of an R-module M is a subset $N \subseteq M$ that is an R-module under the operations of M. Equivalently, N is a subgroup of $(M, +)$ with $rn \in N$ for all $r \in R$ and $n \in N$.

Definition 10.1.8 (Ideal)

An ***ideal*** of R is a submodule of R, viewed as a module over itself: a subgroup I of $(R, +)$ such that $ra \in I$ for all $r \in R$ and $a \in I$. For $a \in R$, the set $aR = \{ra : r \in R\}$ is an ideal, the ***principal ideal*** generated by a.

Example 10.1.9 (Ideals of $\mathbb{Z}$ and $k[x]$)

The ideals of $\mathbb{Z}$ are the subgroups $n\mathbb{Z}$ with $n \geqslant 0$: every subgroup of $\mathbb{Z}$ has this form (Theorem 2.4.5), and each $n\mathbb{Z}$ is closed under multiplication by integers. The ideals of $k[x]$ are $\{0\}$ and the sets $p(x)k[x]$ with $p \neq 0$. Indeed, if $I \neq \{0\}$ is an ideal, let $p \in I$ be a nonzero polynomial of least degree; for $f \in I$, Euclidean division gives $f = qp + r$ with $r = 0$ or $\deg r < \deg p$, and since $r = f - qp \in I$, minimality forces $r = 0$. So $I = p(x)k[x]$.
In both rings every ideal is principal. An integral domain (a commutative ring with $1 \neq 0$ and no zero divisors) in which every ideal is principal is called a ***principal ideal domain***, or ***PID***; so $\mathbb{Z}$ and $k[x]$ are PIDs. In both cases the reason is the Euclidean algorithm, which shows that the ideal $pR + qR$ generated by two elements is the principal ideal generated by $\gcd(p, q)$.

If N is a submodule of M, the quotient group M/N (every subgroup of an abelian group is normal) is an R-module with $r(m + N) = rm + N$. This is well defined: if $m + N = m' + N$, then $rm - rm' = r(m - m') \in N$. The axioms for M/N follow from those for M.

Example 10.1.10 (Quotient modules)

The $\mathbb{Z}$-module $\mathbb{Z}/n\mathbb{Z}$ is the cyclic group $\mathbb{Z}/n$. The $k[x]$-module $k[x]/xk[x]$ is a copy of k: the class of $a_0 + a_1x + \cdots + a_dx^d$ is the class of the constant a_0, and x acts as 0, since $x \cdot a_0 = a_0x \in xk[x]$. We denote this module simply by k.

NOTE If I is an ideal, the quotient R/I is also a ring, with $(a+I)(b+I) = ab+I$. This is well defined: if $a - a' \in I$ and $b - b' \in I$, then $ab - a'b' = a(b-b') + (a-a')b' \in I$.

10.1.3 Module Homomorphisms

Definition 10.1.11 (Module homomorphism)

Let M and N be R-modules. A map $\varphi : M \to N$ is a ***homomorphism of R-modules*** (or is ***R-linear***) if

$$\varphi(u+w) = \varphi(u) + \varphi(w) \quad \text{and} \quad \varphi(au) = a\varphi(u)$$

for all $u, w \in M$ and $a \in R$. The set of these maps is denoted $\mathrm{Hom}_R(M,N)$. An ***isomorphism*** is a bijective homomorphism.

The kernel $\ker\varphi$ and the image $\mathrm{Im}\,\varphi$ of a homomorphism are submodules; for instance, if $\varphi(u) = 0$ then $\varphi(au) = a\varphi(u) = 0$. The First Isomorphism Theorem holds for modules: the map $M/\ker\varphi \to \mathrm{Im}\,\varphi$, $m + \ker\varphi \mapsto \varphi(m)$, is an isomorphism of groups by Theorem 3.3.3, and it is R-linear, so it is an isomorphism of R-modules.

When $R = \mathbb{Z}$, every homomorphism of abelian groups $\varphi : G \to H$ is $\mathbb{Z}$-linear: $\varphi(ng) = n\varphi(g)$ for $n > 0$ by additivity, and $\varphi(-g) = -\varphi(g)$. So abelian groups and $\mathbb{Z}$-modules are the same objects with the same maps; in the language of Chapter 8, the two categories coincide.

NOTE Since R is commutative, $\mathrm{Hom}_R(M,N)$ is itself an R-module, with $(\varphi+\psi)(v) = \varphi(v) + \psi(v)$ and $(a\varphi)(v) = a\varphi(v)$. (The map $a\varphi$ is R-linear because $(a\varphi)(bv) = abv' = b(av')$, where $v' = \varphi(v)$; this is where commutativity is used.) For the modules R^m and R^n, things work as for vector spaces: a homomorphism $\varphi : R^m \to R^n$ is determined by the images $\varphi(e_1), \dots, \varphi(e_m)$ of the standard vectors $e_j = (0, \dots, 1, \dots, 0)$, and these can be arbitrary, so $\mathrm{Hom}_R(R^m, R^n)$ is isomorphic to the module of $n \times m$ matrices with entries in R (the j-th column lists the entries of $\varphi(e_j)$). For other modules, $\mathrm{Hom}_R(M,N)$ can be zero even when M and N are both nonzero.

Example 10.1.12 (A vanishing Hom)

Let $R = k[x]$ and let $k = k[x]/xk[x]$ be the module on which x acts as 0. Then $\mathrm{Hom}_{k[x]}(k, k[x]) = 0$. Indeed, if $\varphi : k \to k[x]$ is $k[x]$-linear, then $x\varphi(1) = \varphi(x \cdot 1) = \varphi(0) = 0$, and a polynomial p with $xp = 0$ is zero; so $\varphi(1) = 0$, and then $\varphi(c) = \varphi(c \cdot 1) = c\varphi(1) = 0$ for every $c \in k$.

10.1.4 Spanning Sets, Independence, and Bases

For a set I, let $R^{(I)}$ be the module of families $(a_i)_{i \in I}$ of elements of R with $a_i = 0$ for all but finitely many i, with componentwise operations; for $I = \{1, \dots, n\}$ this is R^n. A family $(v_i)_{i \in I}$ of elements of an R-module M determines the homomorphism

$$\varphi : R^{(I)} \to M, \qquad (a_i)_{i \in I} \mapsto \sum_{i \in I} a_i v_i,$$

where the sum is finite. The basic notions of linear algebra can all be phrased in terms of φ.

Definition 10.1.13 (Spanning sets, independence, and bases)

Let $(v_i)_{i \in I}$ be a family of elements of an R-module M, and $\varphi : R^{(I)} \to M$ as above.

- The family ***spans*** M (or ***generates*** M) if φ is surjective: every element of M is a finite linear combination $\sum a_i v_i$ with $a_i \in R$. The module M is ***finitely generated*** if it is spanned by a finite family.
- The family is ***linearly independent*** if φ is injective: $\sum a_i v_i = 0$ implies $a_i = 0$ for all i.
- The family is a ***basis*** of M if φ is bijective. A module that has a basis is ***free***.

The span of a family (the image of φ) is a submodule, written $\mathrm{span}(v_i)$ or $\sum Rv_i$.

If $(v_i)_{i \in I}$ is a basis, then $\varphi : R^{(I)} \to M$ is an isomorphism, so the free modules are the modules isomorphic to some $R^{(I)}$. The standard vectors $e_1, \dots, e_n$ form a basis of R^n; in particular R is free with basis $\{1\}$.

When $R = \mathbb{Z}$, the number of elements of a finite basis is determined by the module. Indeed, an isomorphism $\psi : \mathbb{Z}^m \to \mathbb{Z}^n$ maps $2\mathbb{Z}^m$ onto $2\mathbb{Z}^n$ (because $\psi(2x) = 2\psi(x)$), so it induces an isomorphism $\mathbb{Z}^m/2\mathbb{Z}^m \cong \mathbb{Z}^n/2\mathbb{Z}^n$. Reduction mod 2 identifies these quotients with $(\mathbb{Z}/2)^m$ and $(\mathbb{Z}/2)^n$, which have 2^m and 2^n elements, so $m = n$. This number is the ***rank*** of the free $\mathbb{Z}$-module. (Over a field

the rank is the dimension. The rank of a free module with a finite basis is well defined over every nonzero commutative ring, but we will only need $\mathbb{Z}$.)

NOTE Over a field, every finitely generated module is free, since a finite spanning set contains a basis (Chapter 4). Over $\mathbb{Z}$, the finitely generated free modules are the groups $\mathbb{Z}^n$. The module $\mathbb{Z}/n$ is *not* free for $n \geqslant 2$, as we show in the next subsection.

Definition 10.1.14 (Direct sum of modules)

The ***direct sum*** $M_1 \oplus M_2$ of R-modules is the set of pairs (m_1, m_2) with componentwise operations. More generally, for a family of modules $(M_i)_{i\in I}$, the direct sum $\bigoplus_{i\in I} M_i$ consists of the families $(m_i)_{i\in I}$ with $m_i \in M_i$ and only finitely many m_i nonzero. For example, $R^{(I)} = \bigoplus_{i\in I} R$.

The direct sum has the following universal property: homomorphisms $\bigoplus_i M_i \to N$ correspond to families of homomorphisms $\varphi_i : M_i \to N$. A homomorphism out of the direct sum restricts to each summand, and conversely a family (φ_i) gives the homomorphism $(m_i) \mapsto \sum_i \varphi_i(m_i)$, a finite sum. For finitely many summands, $M_1 \oplus \cdots \oplus M_k$ is the same module as the product $M_1 \times \cdots \times M_k$, and for abelian groups we use both notations.

10.1.5 General Facts about Modules: Nothing Is True

Over a field, every vector space has a basis, every linearly independent set extends to a basis, every spanning set contains a basis, and every subspace of a finite-dimensional space is finite-dimensional. Over a general ring, all four statements fail.

Example 10.1.15 (Bases need not exist)

The $\mathbb{Z}$-module $M = \mathbb{Z}/n$, with $n \geqslant 2$, has no basis. If $(v_i)_{i\in I}$ is a nonempty family in M, then $\varphi : \mathbb{Z}^{(I)} \to M$ is not injective: it sends $ne_i \neq 0$ to $nv_i = 0$. So no nonempty family is linearly independent, and the empty family spans only $\{0\} \neq M$.

Example 10.1.16 (Linearly independent sets need not extend to a basis)

In $M = \mathbb{Z}$, the element 2 is linearly independent: $a \cdot 2 = 0$ implies $a = 0$. But no basis of $\mathbb{Z}$ contains 2. A basis cannot contain two elements a and b, because $b \cdot a - a \cdot b = 0$ would be a nontrivial relation (basis elements are

nonzero, since $1 \cdot 0 = 0$). So a basis is a single element a with $a\mathbb{Z} = \mathbb{Z}$, and the only bases of $\mathbb{Z}$ are $\{1\}$ and $\{-1\}$.

Example 10.1.17 (Spanning sets need not contain a basis)

In $M = \mathbb{Z}$, the set $\{4, 5\}$ spans: $1 = 5 \cdot 5 - 6 \cdot 4$, so $n = 5n \cdot 5 - 6n \cdot 4$ for every n. But $\{4, 5\}$ is not linearly independent ($5 \cdot 4 - 4 \cdot 5 = 0$), and neither $\{4\}$ nor $\{5\}$ spans $\mathbb{Z}$. So this spanning set contains no basis.

Example 10.1.18 (Submodules of finitely generated modules need not be finitely generated)

Let $R = k[x_1, x_2, \ldots]$ be the ring of polynomials in infinitely many variables, and $M = R$. Then M is generated by 1, but the ideal $M' = \{f \in R : f(0) = 0\}$ of polynomials with zero constant term is not finitely generated. Indeed, finitely many $f_1, \ldots, f_m \in M'$ involve only finitely many variables, say $x_1, \ldots, x_N$. Setting $x_1 = \cdots = x_N = 0$ is a ring homomorphism $R \to R$ that kills each f_j, hence kills every combination $\sum r_j f_j$; but it does not kill $x_{N+1} \in M'$. So $f_1, \ldots, f_m$ do not span M'.
Over $\mathbb{Z}$ this cannot happen: we will see that every submodule of $\mathbb{Z}^n$ is finitely generated (Lemma 10.2.2). More generally, it cannot happen over *Noetherian* rings, those in which every ideal is finitely generated, such as $k[x_1, \ldots, x_n]$; we will not prove this.

10.1.6 Exact Sequences

Definition 10.1.19 (Exact sequences of modules)

A sequence of homomorphisms of R-modules

$$\cdots \to M_{i-1} \xrightarrow{f_{i-1}} M_i \xrightarrow{f_i} M_{i+1} \to \cdots$$

is ***exact*** at M_i if $\operatorname{Im} f_{i-1} = \ker f_i$, and exact if it is exact at every term that has a map in and a map out, as for groups (Definition 3.5.1). A ***short exact sequence*** $0 \to A \xrightarrow{f} B \xrightarrow{g} C \to 0$ is exact at A, B, and C: f is injective, g is surjective, and $\operatorname{Im} f = \ker g$, so that $C \cong B/f(A)$ by the First Isomorphism Theorem. It ***splits*** if there is a homomorphism $s : C \to B$ with $g \circ s = \mathrm{id}_C$.

Proposition 10.1.20 (Split exact sequences)

Let $0 \to A \xrightarrow{f} B \xrightarrow{g} C \to 0$ be a short exact sequence of R-modules.

(a) If $s : C \to B$ satisfies $g \circ s = \mathrm{id}_C$, then $A \oplus C \to B$, $(a, c) \mapsto f(a) + s(c)$, is an isomorphism.

(b) If C is free, the sequence splits; hence $B \cong A \oplus C$.

Proof. (a) By Corollary 3.5.9, applied to the underlying abelian groups, the map is an isomorphism of groups; it is R-linear because f and s are.

(b) Let $(c_i)_{i \in I}$ be a basis of C. Since g is surjective, choose $b_i \in B$ with $g(b_i) = c_i$. Since $\varphi : R^{(I)} \to C$, $(a_i) \mapsto \sum a_i c_i$, is an isomorphism, there is a homomorphism $s : C \to B$ with $s(c_i) = b_i$, namely $s = \psi \circ \varphi^{-1}$ with $\psi((a_i)) = \sum a_i b_i$. Then $g \circ s$ and id_C agree on the basis, hence everywhere. Part (a) gives $B \cong A \oplus C$. □

10.2 Classification of Finitely Generated Abelian Groups

From now on $R = \mathbb{Z}$. As we saw, $\mathbb{Z}$-modules are abelian groups, submodules are subgroups, and $\mathbb{Z}$-linear maps are group homomorphisms; we write $\mathrm{Hom}(M, N)$ for $\mathrm{Hom}_{\mathbb{Z}}(M, N)$. The goal of this section is the following theorem.

Theorem 10.2.1 (Classification of finitely generated abelian groups)

Every finitely generated abelian group G is isomorphic to a product of cyclic groups

$$G \cong \mathbb{Z}/n_1 \times \mathbb{Z}/n_2 \times \cdots \times \mathbb{Z}/n_k \times \mathbb{Z}^{\ell},$$

where $k, \ell \geqslant 0$ and every $n_i \geqslant 2$. The n_i can be chosen so that $n_1 \mid n_2 \mid \cdots \mid n_k$, or alternatively so that each n_i is a power of a prime. In the second form, ℓ and the prime powers $n_1, \ldots, n_k$ (up to order) are determined by G.

We follow Artin (*Algebra*, §§14.4–14.7). The existence of the decomposition reduces to linear algebra over $\mathbb{Z}$ in two steps: first we write any finitely generated abelian group as $\mathbb{Z}^n / \mathrm{Im}(T)$ for some $T \in \mathrm{Hom}(\mathbb{Z}^m, \mathbb{Z}^n)$; then we choose bases of $\mathbb{Z}^m$ and $\mathbb{Z}^n$ in which T is diagonal.

10.2.1 Step 1: Presenting a Module as a Quotient

Lemma 10.2.2 (Submodules of $\mathbb{Z}^n$ are free)

Every submodule of $\mathbb{Z}^n$ is free, with a basis of at most n elements. In particular, it is finitely generated.

Proof. By induction on n. For $n = 1$, the subgroups of $\mathbb{Z}$ are $\{0\}$ and $a\mathbb{Z}$ with $a \neq 0$ (Theorem 2.4.5), which are free with bases $\varnothing$ and $\{a\}$.

Let $n \geqslant 2$, assume the result for $\mathbb{Z}^{n-1}$, and let $N \subseteq \mathbb{Z}^n$ be a submodule. The projection $\pi : \mathbb{Z}^n \to \mathbb{Z}^{n-1}$, $(a_1, \dots, a_n) \mapsto (a_2, \dots, a_n)$, restricts to a homomorphism $\pi|_N : N \to \mathbb{Z}^{n-1}$.

- The image $\pi(N)$ is a submodule of $\mathbb{Z}^{n-1}$, so by induction it has a basis $g_1, \dots, g_m$ with $m \leqslant n-1$. Choose $f_1, \dots, f_m \in N$ with $\pi(f_j) = g_j$.
- The kernel $\ker(\pi|_N) = N \cap (\mathbb{Z} \times 0 \times \cdots \times 0)$ is a subgroup of $\mathbb{Z} \times 0 \times \cdots \times 0 \cong \mathbb{Z}$, so it has a basis $e_1, \dots, e_k$ with $k \leqslant 1$.

We claim that $e_1, \dots, e_k, f_1, \dots, f_m$ is a basis of N; it has $k + m \leqslant n$ elements.

Spanning. Let $x \in N$. Write $\pi(x) = \sum_j a_j g_j$ with $a_j \in \mathbb{Z}$. Then $\pi\big(x - \sum_j a_j f_j\big) = 0$, so

$$x - \textstyle\sum_j a_j f_j \in \ker(\pi|_N) = \operatorname{span}(e_1, \dots, e_k),$$

and therefore $x \in \operatorname{span}(e_1, \dots, e_k, f_1, \dots, f_m)$.

Independence. Suppose $\sum_i c_i e_i + \sum_j b_j f_j = 0$. Applying π, and using $\pi(e_i) = 0$, gives $\sum_j b_j g_j = 0$; since the g_j are independent, every $b_j = 0$. Then $\sum_i c_i e_i = 0$, and since the e_i are independent, every $c_i = 0$. □

Proposition 10.2.3 (Presentation of finitely generated abelian groups)

If M is a finitely generated $\mathbb{Z}$-module, there exist $m, n \geqslant 0$ and $T \in \operatorname{Hom}(\mathbb{Z}^m, \mathbb{Z}^n)$ such that

$$M \cong \mathbb{Z}^n / \operatorname{Im}(T).$$

Equivalently, there is an exact sequence $\mathbb{Z}^m \xrightarrow{T} \mathbb{Z}^n \to M \to 0$.

Proof. Let $x_1, \dots, x_n$ generate M. Then the homomorphism

$$\varphi : \mathbb{Z}^n \to M, \qquad (a_1, \dots, a_n) \mapsto \textstyle\sum a_i x_i,$$

is surjective. Its kernel N is a submodule of $\mathbb{Z}^n$, so by Lemma 10.2.2 it is generated by finitely many elements $f_1, \dots, f_m$. Let $T : \mathbb{Z}^m \to \mathbb{Z}^n$ be the homomorphism with $T(e_j) = f_j$, that is, $T(b_1, \dots, b_m) = \sum b_j f_j$. Then $\operatorname{Im}(T) = N = \ker \varphi$, so the sequence

$$\mathbb{Z}^m \xrightarrow{T} \mathbb{Z}^n \xrightarrow{\varphi} M \to 0$$

is exact, and the First Isomorphism Theorem gives $M \cong \mathbb{Z}^n/\ker\varphi = \mathbb{Z}^n/\operatorname{Im}(T)$. □

Since submodules of a $\mathbb{Z}$-module are just subgroups, $\mathbb{Z}^n/\operatorname{Im}(T)$ is the quotient group of Chapter 3.

10.2.2 Step 2: Diagonalization over $\mathbb{Z}$

Over a field, we can choose bases in which a linear map has matrix $\left(\begin{smallmatrix} I_r & 0 \\ 0 & 0 \end{smallmatrix}\right)$ (Chapter 6), and the proof divides freely. Over $\mathbb{Z}$ we cannot divide; the Euclidean algorithm takes its place. Throughout this subsection, M and N denote free $\mathbb{Z}$-modules with finite bases, such as $\mathbb{Z}^m$ and $\mathbb{Z}^n$. If $T \in \operatorname{Hom}(M,N)$ and we have chosen bases $(e_1,\dots,e_m)$ of M and $(f_1,\dots,f_n)$ of N, the ***matrix*** $[T]$ of T is the $n \times m$ integer matrix whose j-th column lists the coordinates of $T(e_j)$, as in Chapter 5.

Definition 10.2.4 (Divisibility)

Let M be a free $\mathbb{Z}$-module with a finite basis. The ***divisibility*** $\operatorname{div}(x)$ of a nonzero element $x \in M$ is the largest integer $d \geqslant 1$ such that $x = dy$ for some $y \in M$. An element is ***primitive*** if its divisibility is 1.

If $(e_1,\dots,e_n)$ is any basis of M and $x = \sum a_i e_i$, then $x = dy$ for some y if and only if d divides every a_i (write $y = \sum b_i e_i$ and compare coordinates). Hence

$$\operatorname{div}(x) = \gcd(a_1,\dots,a_n),$$

computed in any basis. In particular $\operatorname{div}(cx) = |c|\operatorname{div}(x)$ for $c \neq 0$, and if one coordinate a_i is nonzero then $\operatorname{div}(x) \leqslant |a_i|$.

Lemma 10.2.5 (Primitive elements and bases)

Let M be a free $\mathbb{Z}$-module with a basis of n elements, and let $v \in M$ be nonzero with divisibility d. Then $v = dw$ for some w that belongs to a basis of M. In particular, v belongs to some basis of M if and only if v is primitive.

Proof. An element e_1 of a basis $(e_1,\dots,e_n)$ is primitive, since its coordinates are $1,0,\dots,0$, whose gcd is 1. This proves "only if."

For the main statement, let $(e_1,\dots,e_n)$ be a basis and $v = a_1e_1 + \cdots + a_ne_n$. Renumbering, we may assume that $a_1 \neq 0$ and $|a_1| = \min\{|a_i| : a_i \neq 0\}$. For each $k \geqslant 2$, Euclidean division gives $a_k = q_k a_1 + r_k$ with $0 \leqslant r_k < |a_1|$. Let $e_1' = e_1 + \sum_{k\geqslant 2} q_k e_k$. Then $(e_1', e_2, \dots, e_n)$ is again a basis (the substitution $e_1 = e_1' - \sum_{k\geqslant 2} q_k e_k$ inverts the change), and

$$v = a_1 e_1' + r_2 e_2 + \cdots + r_n e_n.$$

If some r_k is nonzero, the smallest nonzero absolute value of a coordinate of v has decreased strictly, and we repeat. This cannot go on forever, so after finitely many steps we reach a basis in which v has a single nonzero coordinate: $v = aw$ with w a basis element. Then $d = \operatorname{div}(v) = |a|$, and replacing w by $-w$ if necessary, $v = dw$. If v is primitive, $d = 1$ and $v = w$ belongs to a basis. □

Theorem 10.2.6 (Smith normal form)

Let M and N be free $\mathbb{Z}$-modules with bases of m and n elements (for instance $\mathbb{Z}^m$ and $\mathbb{Z}^n$), and let $T \in \operatorname{Hom}(M, N)$. There exist bases $(e_1, \dots, e_m)$ of M and $(f_1, \dots, f_n)$ of N, an integer r with $0 \leqslant r \leqslant \min(m, n)$, and positive integers $d_1 \mid d_2 \mid \cdots \mid d_r$ such that

$$T(e_i) = \begin{cases} d_i f_i & \text{if } 1 \leqslant i \leqslant r, \\ 0 & \text{if } r < i \leqslant m. \end{cases}$$

In matrix form,

$$[T] = \left(\begin{array}{ccc|c} d_1 & & & \\ & \ddots & & 0 \\ & & d_r & \\ \hline & 0 & & 0 \end{array}\right).$$

Moreover, if $T \neq 0$, then d_1 is the greatest common divisor of the entries of the matrix of T in any bases of M and N.

Proof. By induction on m. If $T = 0$ (in particular, if $m = 0$), take $r = 0$ and any bases. So let $T \neq 0$, and assume the result for homomorphisms out of free modules with bases of $m - 1$ elements.

Choosing e_1 and f_1. Let

$$d_1 = \min\{\operatorname{div}(T(x)) : x \in M,\ T(x) \neq 0\},$$

and choose $e_1 \in M$ with $\operatorname{div}(T(e_1)) = d_1$. Then e_1 is primitive: if $e_1 = cx$ with $c \geqslant 1$, then $T(e_1) = cT(x)$, so $T(x) \neq 0$ and $\operatorname{div}(T(x)) = d_1/c$, and minimality forces $c = 1$. By Lemma 10.2.5, e_1 belongs to a basis $(e_1, \dots, e_m)$ of M, and $T(e_1) = d_1 f_1$ for some f_1 belonging to a basis $(f_1, \dots, f_n)$ of N.

Clearing the first row. For $j \geqslant 2$, let a_j be the f_1-coordinate of $T(e_j)$, and write $a_j = q_j d_1 + r_j$ with $0 \leqslant r_j < d_1$. Replace e_j by $e'_j = e_j - q_j e_1$; then $(e_1, e'_2, \dots, e'_m)$ is still a basis, and the f_1-coordinate of $T(e'_j) = T(e_j) - q_j d_1 f_1$ is r_j. If $r_j \neq 0$, then $T(e'_j) \neq 0$ and $\operatorname{div}(T(e'_j)) \leqslant r_j < d_1$, contradicting the choice of d_1. So every $r_j = 0$, and T maps $M' = \operatorname{span}(e'_2, \dots, e'_m)$ into $N' = \operatorname{span}(f_2, \dots, f_n)$. In these

bases,

$$[T] = \left(\begin{array}{c|c} d_1 & 0 \\ \hline 0 & [T'] \end{array}\right),$$

where $T' : M' \to N'$ is the restriction of T.

Divisibility by d_1. Let $y \in M'$. Then $T(e_1 + y) = d_1 f_1 + T(y)$ is nonzero, and its divisibility is the gcd of d_1 and the coordinates of $T(y) \in N'$, which is at most d_1. By minimality it equals d_1, so d_1 divides every coordinate of $T(y)$. Thus $T(M') \subseteq d_1 N'$.

Induction. The modules M' and N' are free with bases of $m-1$ and $n-1$ elements. By the induction hypothesis there are bases $(e_2'', \dots, e_m'')$ of M' and $(f_2'', \dots, f_n'')$ of N', an integer r with $1 \leqslant r \leqslant \min(m, n)$, and positive integers $d_2 \mid \cdots \mid d_r$, such that $T(e_i'') = d_i f_i''$ for $2 \leqslant i \leqslant r$ and $T(e_i'') = 0$ for $i > r$. Since $M = \mathbb{Z}e_1 \oplus M'$ and $N = \mathbb{Z}f_1 \oplus N'$, the families $(e_1, e_2'', \dots, e_m'')$ and $(f_1, f_2'', \dots, f_n'')$ are bases of M and N, and in them T has the required form. If $r \geqslant 2$, then $d_2 f_2'' = T(e_2'') \in d_1 N'$; the coordinates of $d_2 f_2''$ in the basis $(f_2'', \dots, f_n'')$ are $d_2, 0, \dots, 0$, so $d_1 \mid d_2$.

The gcd of the entries. Let A be the matrix of T in any bases (e_j^*) of M and (f_i^*) of N, and let g be the gcd of its entries. In the bases just constructed, $T(\sum x_i e_i) = \sum_{i \leqslant r} x_i d_i f_i$, and every d_i is a multiple of d_1, so $T(M) \subseteq d_1 N$. In particular each $T(e_j^*)$ has all its coordinates divisible by d_1; these coordinates are the entries of A, so $d_1 \mid g$. Conversely, the coordinates of any $T(x)$ in the basis (f_i^*) are integer combinations of the entries of A, hence divisible by g; so g divides $\operatorname{div}(T(e_1)) = d_1$. Therefore $g = d_1$. □

NOTE The Smith normal form is the integer analogue of row reduction. Replacing a basis vector e_j of M by $e_j + c e_k$ (with $k \neq j$ and $c \in \mathbb{Z}$), swapping two basis vectors, or replacing e_j by $-e_j$ performs the corresponding operation on the columns of $[T]$: adding c times column k to column j, swapping two columns, negating a column. Likewise, every integer row operation (adding an integer multiple of one row to another, swapping two rows, negating a row) is the effect of such a change of basis of N. Since these operations can be undone over $\mathbb{Z}$, the new families are again bases. So the theorem says that integer row and column operations reduce every integer matrix to the diagonal form above. Over a field, Gaussian elimination reaches a diagonal matrix with entries 1 and 0; over $\mathbb{Z}$, the entries d_i record what cannot be removed without dividing.

10.2.3 Step 3: Existence of the Decomposition

Lemma 10.2.7 (Chinese Remainder Theorem for cyclic groups)
If $\gcd(m,n)=1$, then $\mathbb{Z}/mn \cong \mathbb{Z}/m \times \mathbb{Z}/n$. Consequently, if $n = p_1^{a_1}\cdots p_s^{a_s}$ with distinct primes p_i, then $\mathbb{Z}/n \cong \mathbb{Z}/p_1^{a_1} \times \cdots \times \mathbb{Z}/p_s^{a_s}$.

Proof. In $\mathbb{Z}/m \times \mathbb{Z}/n$, the multiple $c\cdot(1,1) = (c \bmod m,\, c \bmod n)$ is zero if and only if $m \mid c$ and $n \mid c$, that is, if and only if $mn \mid c$, because m and n are coprime. So $(1,1)$ has order mn, which is the order of the group. The group is therefore cyclic, generated by $(1,1)$, and isomorphic to $\mathbb{Z}/mn$. The second statement follows by induction on s, since $p_1^{a_1}\cdots p_{s-1}^{a_{s-1}}$ and $p_s^{a_s}$ are coprime. □

Proof of existence in Theorem 10.2.1. Let G be a finitely generated abelian group. By Proposition 10.2.3, $G \cong \mathbb{Z}^n/\operatorname{Im}(T)$ for some $T \in \operatorname{Hom}(\mathbb{Z}^m,\mathbb{Z}^n)$. By Theorem 10.2.6, there are bases (e_i) of $\mathbb{Z}^m$ and (f_j) of $\mathbb{Z}^n$ with $T(e_i) = d_i f_i$ for $i \leqslant r$ and $T(e_i) = 0$ for $i > r$, where $d_1 \mid \cdots \mid d_r$. Then $T\big(\sum x_i e_i\big) = \sum_{i\leqslant r} x_i d_i f_i$, so

$$\operatorname{Im}(T) = \{a_1 f_1 + \cdots + a_n f_n : d_i \mid a_i \text{ for } i \leqslant r,\ a_i = 0 \text{ for } i > r\}.$$

The surjective homomorphism $\mathbb{Z}^n \to \mathbb{Z}/d_1 \times \cdots \times \mathbb{Z}/d_r \times \mathbb{Z}^{n-r}$ given by

$$\sum a_i f_i \mapsto (a_1 \bmod d_1, \ldots, a_r \bmod d_r, a_{r+1}, \ldots, a_n)$$

has kernel exactly $\operatorname{Im}(T)$, so by the First Isomorphism Theorem

$$G \cong \mathbb{Z}^n/\operatorname{Im}(T) \cong \mathbb{Z}/d_1 \times \cdots \times \mathbb{Z}/d_r \times \mathbb{Z}^{n-r}.$$

The factors $\mathbb{Z}/d_i$ with $d_i = 1$ are trivial and can be dropped; since $d_1 \mid d_2 \mid \cdots$, they are the first few. Calling the remaining d_i (all $\geqslant 2$) $n_1, \ldots, n_k$ and setting $\ell = n - r$ gives

$$G \cong \mathbb{Z}/n_1 \times \cdots \times \mathbb{Z}/n_k \times \mathbb{Z}^{\ell}, \qquad n_1 \mid n_2 \mid \cdots \mid n_k.$$

Finally, Lemma 10.2.7 splits each $\mathbb{Z}/n_i$ into a product of cyclic groups of prime power order, which gives the second form. □

10.2.4 Uniqueness of the Decomposition

To see that the prime power decomposition is determined by G, we read its pieces off subgroups that any isomorphism preserves. The ***torsion subgroup*** of an abelian group G is

$$G_{\mathrm{tors}} = \{g \in G : ng = 0 \text{ for some } n \geqslant 1\},$$

the set of elements of finite order; it is a subgroup, because $ng = 0$ and $mh = 0$ imply $nm(g+h) = 0$. For a prime p and $j \geqslant 0$, let $G[p^j] = \{g \in G : p^j g = 0\}$, also a subgroup. An isomorphism $G \to G'$ maps G_{tors} onto G'_{tors} and $G[p^j]$ onto $G'[p^j]$.

Proposition 10.2.8 (Uniqueness)

Suppose that

$$G \cong \mathbb{Z}^\ell \times \mathbb{Z}/p_1^{a_1} \times \cdots \times \mathbb{Z}/p_k^{a_k}$$

with primes p_i (not necessarily distinct) and exponents $a_i \geqslant 1$. Then $G/G_{\text{tors}} \cong \mathbb{Z}^\ell$, and for every prime p and every $j \geqslant 1$,

$$\left|G[p^j]\right| / \left|G[p^{j-1}]\right| = p^{\,c_j(p)}, \qquad c_j(p) = \#\{i : p_i = p \text{ and } a_i \geqslant j\}.$$

Consequently ℓ and the prime powers $p_1^{a_1}, \ldots, p_k^{a_k}$, up to order, are determined by G.

Proof. Since isomorphisms preserve torsion subgroups and the subgroups $G[p^j]$, we may assume that G equals the product, and we write its elements as (x, y) with $x \in \mathbb{Z}^\ell$ and $y \in A = \mathbb{Z}/p_1^{a_1} \times \cdots \times \mathbb{Z}/p_k^{a_k}$.

Every element of the finite group A has finite order, and no nonzero element of $\mathbb{Z}^\ell$ does. So $G_{\text{tors}} = 0 \times A$, which is the kernel of the projection $G \to \mathbb{Z}^\ell$; hence $G/G_{\text{tors}} \cong \mathbb{Z}^\ell$. Since the rank of a free $\mathbb{Z}$-module is well defined, ℓ is determined by G.

For the same reason $G[p^j] = 0 \times A[p^j]$, and $A[p^j]$ is the product of the subgroups $C[p^j]$ of its factors $C = \mathbb{Z}/q^a$. If $q \neq p$ and $p^j x \equiv 0 \pmod{q^a}$, then $q^a \mid x$ because p and q are coprime; so $C[p^j] = 0$. If $q = p$, then $p^a \mid p^j x$ if and only if $p^{a-\min(a,j)} \mid x$, so $C[p^j]$ has $p^{\min(a,j)}$ elements. Hence

$$\left|G[p^j]\right| = p^{\sum_{i:p_i=p} \min(a_i, j)}.$$

Since $\min(a,j) - \min(a,j-1)$ is 1 if $a \geqslant j$ and 0 otherwise, the quotient $|G[p^j]|/|G[p^{j-1}]|$ is $p^{c_j(p)}$. These numbers depend only on G, and the number of indices i with $p_i^{a_i} = p^j$ is $c_j(p) - c_{j+1}(p)$. So the list of prime powers is determined by G. □

This completes the proof of Theorem 10.2.1.

NOTE The number ℓ is the ***free rank*** (or ***Betti number***) of G, and the prime powers $n_1, \ldots, n_k$ in the second form are its ***elementary divisors***. The integers $n_1 \mid \cdots \mid n_k$ of the first form are called the ***invariant factors***; they are determined by the elementary divisors (the last invariant factor is the

product of the largest power of each prime, and so on), so they too depend only on G. The free rank counts the copies of $\mathbb{Z}$, while the elementary divisors describe the torsion subgroup $G_{\text{tors}} \cong \mathbb{Z}/n_1 \times \cdots \times \mathbb{Z}/n_k$.

Example 10.2.9 (Abelian groups of small order)

By the theorem, the abelian groups of a given finite order, up to isomorphism, correspond to the ways of writing the order as a product of prime powers, one factor $\mathbb{Z}/p^a$ for each.

- Order 4: $\mathbb{Z}/4$ or $\mathbb{Z}/2 \times \mathbb{Z}/2$. (Directly: the first has an element of order 4, the second does not.)
- Order 6: only $\mathbb{Z}/6 \cong \mathbb{Z}/2 \times \mathbb{Z}/3$.
- Order 8: $\mathbb{Z}/8$, $\mathbb{Z}/4 \times \mathbb{Z}/2$, or $\mathbb{Z}/2 \times \mathbb{Z}/2 \times \mathbb{Z}/2$.
- Order 12: $\mathbb{Z}/12 \cong \mathbb{Z}/4 \times \mathbb{Z}/3$, or $\mathbb{Z}/2 \times \mathbb{Z}/6 \cong \mathbb{Z}/2 \times \mathbb{Z}/2 \times \mathbb{Z}/3$.

Exercise 10.2.10 (Modules and finitely generated abelian groups)

(1) Classify the abelian groups of order 36 up to isomorphism.

(2) Let $T : \mathbb{Z}^2 \to \mathbb{Z}^2$ be the homomorphism with matrix $A = \begin{pmatrix} 6 & 4 \\ 3 & 5 \end{pmatrix}$. Compute the Smith normal form of T and identify the group $\mathbb{Z}^2/\operatorname{Im}(T)$.

(3) Prove that a finitely generated abelian group is free (that is, isomorphic to $\mathbb{Z}^n$ for some n) if and only if it is torsion-free (its only element of finite order is 0).

(4) Let G be an abelian group of order p^n, where p is prime. Show that $G \cong \mathbb{Z}/p^{a_1} \times \cdots \times \mathbb{Z}/p^{a_k}$ with $a_i \geqslant 1$ and $a_1 + \cdots + a_k = n$. How many such groups are there, up to isomorphism?

(5) Show that $\operatorname{Hom}(\mathbb{Z}/m, \mathbb{Z}/n) \cong \mathbb{Z}/\gcd(m, n)$ as abelian groups. (Hint: a homomorphism is determined by the image of 1.)

Solution. **(1)** We have $36 = 2^2 \cdot 3^2$. The factors of 2-power order are $\mathbb{Z}/4$ or $\mathbb{Z}/2 \times \mathbb{Z}/2$, and those of 3-power order are $\mathbb{Z}/9$ or $\mathbb{Z}/3 \times \mathbb{Z}/3$. Taking all combinations:

$$\mathbb{Z}/4 \times \mathbb{Z}/9, \quad \mathbb{Z}/4 \times \mathbb{Z}/3 \times \mathbb{Z}/3, \quad \mathbb{Z}/2 \times \mathbb{Z}/2 \times \mathbb{Z}/9, \quad \mathbb{Z}/2 \times \mathbb{Z}/2 \times \mathbb{Z}/3 \times \mathbb{Z}/3.$$

By the uniqueness part of Theorem 10.2.1, these are pairwise non-isomorphic, so there are exactly 4 abelian groups of order 36. In invariant factor form (Lemma 10.2.7) they are $\mathbb{Z}/36$, $\mathbb{Z}/3 \times \mathbb{Z}/12$, $\mathbb{Z}/2 \times \mathbb{Z}/18$, and $\mathbb{Z}/6 \times \mathbb{Z}/6$.

(2) The gcd of the entries of A is 1, so $d_1 = 1$ by Theorem 10.2.6. To find the bases, we apply integer row and column operations (see the note after Theorem 10.2.6):

$$\begin{pmatrix} 6 & 4 \\ 3 & 5 \end{pmatrix} \xrightarrow{R_1 - 2R_2} \begin{pmatrix} 0 & -6 \\ 3 & 5 \end{pmatrix} \xrightarrow{R_1 \leftrightarrow R_2} \begin{pmatrix} 3 & 5 \\ 0 & -6 \end{pmatrix} \xrightarrow{C_2 - C_1} \begin{pmatrix} 3 & 2 \\ 0 & -6 \end{pmatrix}$$
$$\xrightarrow{C_1 - C_2} \begin{pmatrix} 1 & 2 \\ 6 & -6 \end{pmatrix} \xrightarrow{C_2 - 2C_1} \begin{pmatrix} 1 & 0 \\ 6 & -18 \end{pmatrix} \xrightarrow{R_2 - 6R_1} \begin{pmatrix} 1 & 0 \\ 0 & -18 \end{pmatrix} \xrightarrow{-R_2} \begin{pmatrix} 1 & 0 \\ 0 & 18 \end{pmatrix}.$$

Here $R_1 - 2R_2$ means "subtract twice row 2 from row 1," and so on. The Smith normal form is $\operatorname{diag}(1, 18)$, so $\mathbb{Z}^2/\operatorname{Im}(T) \cong \mathbb{Z}/1 \times \mathbb{Z}/18 \cong \mathbb{Z}/18 \cong \mathbb{Z}/2 \times \mathbb{Z}/9$. As a check, $d_1 d_2 = 18 = |\det A| = |30 - 12|$, as it must be: adding a multiple of one row or column to another does not change the determinant, and swapping or negating rows or columns changes only its sign.

(3) If $G \cong \mathbb{Z}^n$, then G is torsion-free, since $cx = 0$ with $c \geqslant 1$ and $x \in \mathbb{Z}^n$ forces $x = 0$.

Conversely, let G be finitely generated and torsion-free, and write $G \cong \mathbb{Z}/n_1 \times \cdots \times \mathbb{Z}/n_k \times \mathbb{Z}^\ell$ by Theorem 10.2.1. If $k \geqslant 1$, the element $(1, 0, \ldots, 0)$ is nonzero of finite order n_1. So $k = 0$ and $G \cong \mathbb{Z}^\ell$.

(4) A finite group is finitely generated, so by Theorem 10.2.1, $G \cong \mathbb{Z}^\ell \times \mathbb{Z}/q_1^{a_1} \times \cdots \times \mathbb{Z}/q_k^{a_k}$ with primes q_i and $a_i \geqslant 1$. Since G is finite, $\ell = 0$. Then $p^n = |G| = q_1^{a_1} \cdots q_k^{a_k}$, and by unique factorization every $q_i = p$ and $a_1 + \cdots + a_k = n$.

Conversely, every partition $n = a_1 + \cdots + a_k$ (with $a_1 \geqslant \cdots \geqslant a_k \geqslant 1$) gives a group $\mathbb{Z}/p^{a_1} \times \cdots \times \mathbb{Z}/p^{a_k}$ of order p^n, and by Proposition 10.2.8, different partitions give non-isomorphic groups. So the number of abelian groups of order p^n is the number $p(n)$ of partitions of n, independently of p. For example, $p(4) = 5$: the partitions 4, $3+1$, $2+2$, $2+1+1$, $1+1+1+1$ give five groups of order p^4.

(5) Let $d = \gcd(m, n)$. A homomorphism $\varphi : \mathbb{Z}/m \to \mathbb{Z}/n$ is determined by $x = \varphi(1)$, since $\varphi(a \bmod m) = ax$, and x must satisfy $mx = \varphi(m \cdot 1) = \varphi(0) = 0$. Conversely, if $x \in \mathbb{Z}/n$ satisfies $mx = 0$, then $a \bmod m \mapsto ax$ is well defined (if $m \mid a - a'$ then $ax - a'x = 0$) and is a homomorphism. Since $(\varphi + \psi)(1) = \varphi(1) + \psi(1)$, we get an isomorphism of groups

$$\operatorname{Hom}(\mathbb{Z}/m, \mathbb{Z}/n) \cong \{x \in \mathbb{Z}/n : mx = 0\}.$$

For $x = c \bmod n$, the condition $mx = 0$ says $n \mid mc$, that is, $\frac{n}{d} \mid \frac{m}{d}c$, and since $\frac{n}{d}$ and $\frac{m}{d}$ are coprime, this means $\frac{n}{d} \mid c$. So the subgroup is generated

by $\frac{n}{d}$ mod n, which has order d, and $\mathrm{Hom}(\mathbb{Z}/m, \mathbb{Z}/n) \cong \mathbb{Z}/d$. In particular, if $\gcd(m, n) = 1$ there are no nonzero homomorphisms $\mathbb{Z}/m \to \mathbb{Z}/n$.

CHAPTER 11

Group Actions, Conjugacy, and the Sylow Theorems

We return to group theory with the language of group actions from Chapter 3 in hand. Understanding which sets a group G acts on, and how, reveals the internal structure of G itself. In this chapter we count orbits with Burnside's lemma; let G act on itself by conjugation, which leads to the class equation and to the basic structure of p-groups; classify the finite groups of rotations of $\mathbb{R}^3$; study conjugacy in the symmetric and alternating groups, proving that A_n is simple for $n \geqslant 5$; and prove the Sylow theorems. Together with semidirect products, these results classify the groups of several small orders.

11.1 Orbits, Stabilizers, and Burnside's Lemma

11.1.1 Group Actions

We recall the language of Section 3.2. An *action* of a group G on a set S (Definition 3.2.1) is a map $G \times S \to S$, $(g, s) \mapsto g \cdot s$, with $e \cdot s = s$ and $(gh) \cdot s = g \cdot (h \cdot s)$ for all $g, h \in G$ and $s \in S$; equivalently, it is a homomorphism $\rho : G \to \mathrm{Sym}(S)$, $\rho(g)(s) = g \cdot s$. The action is *faithful* if ρ is injective; in general it comes from a faithful action of $G/\ker\rho$ on S. The *orbit* of s is $G \cdot s = \{g \cdot s : g \in G\}$. The orbits are the equivalence classes of the relation $s \sim t \iff t = g \cdot s$ for some $g \in G$, so they partition S. The action is *transitive* if there is only one orbit; restricting the action to each orbit, we see that every action is a disjoint union of transitive actions. The *stabilizer* $\mathrm{Stab}(s) = \{g \in G : g \cdot s = s\}$ is a subgroup of G.

The basic example is the action of G on the set G/H of left cosets of a subgroup H by left multiplication, $g \cdot aH = gaH$. It is transitive (ba^{-1} maps aH to bH), the stabilizer of the coset H is H itself, and the stabilizer of aH is aHa^{-1}. The Orbit–Stabilizer Theorem (Theorem 3.2.9) says that every transitive action looks like this one: if $s \in S$ and $H = \mathrm{Stab}(s)$, then

$$\varepsilon : G/H \to G \cdot s, \qquad aH \mapsto a \cdot s$$

is a bijection. It is moreover *equivariant*, $\varepsilon(g \cdot aH) = (ga) \cdot s = g \cdot \varepsilon(aH)$, so it identifies the action of G on the orbit of s with the action of G on $G/\operatorname{Stab}(s)$. When G is finite this gives

$$|G \cdot s| = [G : \operatorname{Stab}(s)] = \frac{|G|}{|\operatorname{Stab}(s)|},$$

so the size of every orbit divides $|G|$; and when S is finite, $|S|$ is the sum of the sizes of the orbits.

Proposition 11.1.1 (Conjugate stabilizers)

Let G act on S. If $s' = g \cdot s$, then $\operatorname{Stab}(s') = g \operatorname{Stab}(s) g^{-1}$. In particular, points in the same orbit have conjugate stabilizers, which have the same order.

Proof. If $h \in \operatorname{Stab}(s)$, then $(ghg^{-1}) \cdot s' = (ghg^{-1}g) \cdot s = g \cdot (h \cdot s) = g \cdot s = s'$, so $g \operatorname{Stab}(s) g^{-1} \subseteq \operatorname{Stab}(s')$. Applying this to $s = g^{-1} \cdot s'$ gives $g^{-1} \operatorname{Stab}(s') g \subseteq \operatorname{Stab}(s)$, that is, $\operatorname{Stab}(s') \subseteq g \operatorname{Stab}(s) g^{-1}$. □

Example 11.1.2 (Rotations of the tetrahedron)

Let G be the group of rotations of $\mathbb{R}^3$ that map a regular tetrahedron centered at the origin to itself, acting on the set S of its four faces. Any face can be moved to any other, so the action is transitive and $|G \cdot s| = 4$. The stabilizer of a face consists of the rotations by 0, $2\pi/3$ and $4\pi/3$ about the axis through the center of that face and the opposite vertex, so $|\operatorname{Stab}(s)| = 3$. Hence $|G| = 4 \cdot 3 = 12$: the identity, the eight rotations by $\pm 2\pi/3$ about the four axes through a vertex and the center of the opposite face, and the three rotations by π about the three axes through the midpoints of opposite edges. Numbering the vertices $1, 2, 3, 4$, the rotations of order 3 permute the vertices by 3-cycles, and the three half-turns by $(1\,2)(3\,4)$, $(1\,3)(2\,4)$ and $(1\,4)(2\,3)$. We show in Section 11.3 that $G \cong A_4$.

11.1.2 Burnside's Lemma

For $g \in G$, the set of ***fixed points*** of g is $S^g = \{s \in S : g \cdot s = s\}$. Burnside's lemma counts the orbits of an action in terms of fixed points.

Theorem 11.1.3 (Burnside's Lemma (Cauchy–Frobenius))

Let a finite group G act on a finite set S. The number of orbits is

$$\#\text{orbits} = \frac{1}{|G|} \sum_{g \in G} |S^g|.$$

In words: the number of orbits is the average number of fixed points of an element of G.

Proof. Count the set $\Sigma = \{(g, s) \in G \times S : g \cdot s = s\}$ in two ways. Summing over G gives $|\Sigma| = \sum_{g \in G} |S^g|$. Summing over S gives $|\Sigma| = \sum_{s \in S} |\mathrm{Stab}(s)|$. Group this sum by orbits: for s in an orbit $\mathcal{O}$, $|\mathrm{Stab}(s)| = |G|/|\mathcal{O}|$ by the Orbit–Stabilizer Theorem, so

$$|\Sigma| = \sum_{\mathcal{O}} \sum_{s \in \mathcal{O}} \frac{|G|}{|\mathcal{O}|} = \sum_{\mathcal{O}} |G| = |G| \cdot (\#\text{orbits}).$$

Equating the two counts gives the formula. □

Example 11.1.4 (Colorings of a tetrahedron)

In how many ways can we color the faces of a regular tetrahedron with 3 colors, up to rotation? Here S is the set of colorings of the 4 faces, $|S| = 3^4 = 81$, and G is the rotation group of Example 11.1.2, with 12 elements.

- The identity fixes all 81 colorings.
- Each of the 8 rotations of order 3 fixes one face and permutes the other three cyclically, so a fixed coloring gives those three faces one color: $|S^g| = 3 \cdot 3 = 9$.
- Each of the 3 half-turns swaps the faces in two pairs, so a fixed coloring is constant on each pair: $|S^g| = 3 \cdot 3 = 9$.

So the number of orbits is $\frac{1}{12}(81 + 8 \cdot 9 + 3 \cdot 9) = \frac{180}{12} = 15$.

NOTE The same method handles colorings that are harder to count by hand. To color the six edges of a tetrahedron with 3 colors: the identity fixes 3^6 colorings; a rotation of order 3 permutes the edges in two 3-cycles (the three edges at its vertex and the three edges of the opposite face), so it fixes 3^2; a half-turn maps the two edges its axis passes through to themselves and swaps the other four in two pairs, so it fixes 3^4. The number of colorings up to rotation is $\frac{1}{12}(3^6 + 8 \cdot 3^2 + 3 \cdot 3^4) = \frac{1}{12}(729 + 72 + 243) = 87$.

11.2 Conjugation and the Class Equation

11.2.1 Actions of a Group on Itself

A group acts on itself in two basic ways. Under left multiplication, $g \cdot h = gh$, the action is transitive, every stabilizer is trivial, and no $g \neq e$ has a fixed point. In particular the action is faithful, which gives Cayley's theorem (Theorem 3.2.14): a finite group G is isomorphic to a subgroup of S_n, where $n = |G|$. This embedding tells us little about the structure of G. The second action is far more informative.

The group G also acts on itself by conjugation, $g \cdot h = ghg^{-1}$. Each map $h \mapsto ghg^{-1}$ is an automorphism of G, and $g \mapsto (h \mapsto ghg^{-1})$ is a homomorphism $G \to \mathrm{Aut}(G) \subseteq \mathrm{Sym}(G)$, so this is indeed an action. Recall from Definition 3.2.12 that its orbits are the conjugacy classes $C_h = \{ghg^{-1} : g \in G\}$, and that the stabilizer of h is its *centralizer* $Z(h) = \{g \in G : gh = hg\}$, the subgroup of elements that commute with h; we write $Z_G(h)$ when the group must be named. The kernel of the action is the center $Z(G) = \bigcap_{h \in G} Z(h)$ (Definition 3.1.13).

> NOTE The conjugation action is trivial (every element is fixed) if and only if G is abelian, and it is faithful if and only if $Z(G) = \{e\}$. A subgroup $N \leqslant G$ is normal if and only if it is a union of conjugacy classes (Proposition 3.2.13); it contains the class $\{e\}$, and by Lagrange's theorem its order divides $|G|$. We use this repeatedly to find normal subgroups.

11.2.2 The Class Equation

By Proposition 3.2.13, the conjugacy class of h has $|C_h| = [G : Z(h)]$ elements, a divisor of $|G|$ when G is finite, and $|C_h| = 1$ if and only if $h \in Z(G)$.

Theorem 11.2.1 (The Class Equation)

Let G be a finite group, and let $h_1, \dots, h_r$ be representatives of the conjugacy classes of G with more than one element. Then

$$|G| = |Z(G)| + \sum_{i=1}^{r} [G : Z(h_i)],$$

where each $[G : Z(h_i)] = |C_{h_i}|$ divides $|G|$ and is greater than 1.

Proof. The conjugacy classes partition G. The classes with one element are the sets $\{z\}$ with $z \in Z(G)$, and together they contain $|Z(G)|$ elements. The other classes are $C_{h_1}, \dots, C_{h_r}$, and $|C_{h_i}| = [G : Z(h_i)]$ by the Orbit–Stabilizer Theorem. □

Example 11.2.2 (The class equation of S_3)

By Theorem 3.4.6, the conjugacy classes of S_3 are $\{e\}$, the three transpositions, and the two 3-cycles. So $Z(S_3) = \{e\}$ and the class equation is $6 = 1 + 3 + 2$. We compute more class equations in Section 11.4.

11.2.3 p-Groups

Definition 11.2.3 (p-group)

Let p be a prime. A finite group G is a ***p-group*** if $|G| = p^k$ for some $k \geqslant 0$.

Theorem 11.2.4 (p-groups have nontrivial center)

If G is a p-group and $G \neq \{e\}$, then p divides $|Z(G)|$. In particular $|Z(G)| \geqslant p$.

Proof. Let $|G| = p^k$ with $k \geqslant 1$. In the class equation each $[G : Z(h_i)]$ divides p^k and is greater than 1, so it is divisible by p. Hence $|Z(G)| = |G| - \sum_i [G : Z(h_i)]$ is divisible by p. Since $e \in Z(G)$, $|Z(G)| \geqslant 1$, and therefore $|Z(G)| \geqslant p$. □

Corollary 11.2.5 (Groups of order p^2)

If $|G| = p^2$ with p prime, then G is abelian, and $G \cong \mathbb{Z}/p^2$ or $G \cong \mathbb{Z}/p \times \mathbb{Z}/p$.

Proof. By Theorem 11.2.4, p divides $|Z(G)|$, and $|Z(G)|$ divides p^2 by Lagrange's theorem, so $|Z(G)|$ is p or p^2. Suppose $|Z(G)| = p$, and pick $g \notin Z(G)$. The centralizer $Z(g)$ contains $Z(G)$ and g, so $|Z(g)| > p$; as a subgroup of G its order divides p^2, so $Z(g) = G$. Then g commutes with every element, so $g \in Z(G)$, a contradiction. Hence $Z(G) = G$ and G is abelian. By the classification of finitely generated abelian groups (Theorem 10.2.1), G is a product of cyclic groups whose orders multiply to p^2, so $G \cong \mathbb{Z}/p^2$ or $G \cong \mathbb{Z}/p \times \mathbb{Z}/p$. These are not isomorphic, since only the first has an element of order p^2. □

The next lemma is elementary, and we use it several times in this chapter.

Lemma 11.2.6 (Subgroups of index 2)

Let H be a subgroup of index 2 in a group G. Then H is normal in G, and $g^2 \in H$ for every $g \in G$. In particular, H contains every element of G of odd order.

Proof. The subgroup H is normal by Proposition 3.1.11. The quotient G/H has order 2, so $(gH)^2 = H$, that is, $g^2 \in H$. If g has odd order $2k+1$, then $g = g^{2k+2} = (g^{k+1})^2 \in H$. □

Proposition 11.2.7 (Groups of order 8)

Up to isomorphism there are exactly five groups of order 8: the abelian groups $\mathbb{Z}/8$, $\mathbb{Z}/4 \times \mathbb{Z}/2$ and $(\mathbb{Z}/2)^3$, the dihedral group D_4 of symmetries of the square, and the ***quaternion group*** $Q_8 = \{\pm 1, \pm i, \pm j, \pm k\}$, in which -1 commutes with every element, $(-1)^2 = 1$, multiplication by -1 changes signs, and

$$i^2 = j^2 = k^2 = -1, \qquad ij = -ji = k, \qquad jk = -kj = i, \qquad ki = -ik = j.$$

Proof. First, Q_8 is a group. In $GL_2(\mathbb{C})$, the matrices $A = \left(\begin{smallmatrix} i & 0 \\ 0 & -i \end{smallmatrix}\right)$ and $B = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ satisfy $A^2 = B^2 = (AB)^2 = -I$, $BA = -AB$, $B(AB) = A$ and $(AB)A = B$. So the eight distinct matrices $\pm I, \pm A, \pm B, \pm AB$ multiply by the rules above, with $1, i, j, k$ replaced by I, A, B, AB; they form a subgroup of $GL_2(\mathbb{C})$, and Q_8 is a group isomorphic to it.

The three abelian groups listed are pairwise non-isomorphic (the largest orders of their elements are 8, 4 and 2), and by Theorem 10.2.1 every abelian group of order 8 is one of them. The groups D_4 and Q_8 are not abelian, and they are not isomorphic: D_4 has five elements of order 2 (the four reflections and the rotation by π), while -1 is the only element of order 2 in Q_8.

Now let G be a non-abelian group of order 8. No element of G has order 8, since G is not cyclic. If every element satisfied $g^2 = e$, then $ab = (ab)^{-1} = b^{-1}a^{-1} = ba$ for all a, b, and G would be abelian. So some $a \in G$ has order 4. The subgroup $N = \langle a \rangle$ has index 2, so it is normal by Lemma 11.2.6. Fix $b \notin N$. Then $G = N \sqcup Nb$, and every element of G can be written uniquely as $a^i b^k$ with $0 \leqslant i \leqslant 3$ and $k \in \{0, 1\}$.

Since N is normal, $bab^{-1} \in N$, and it has order 4, so $bab^{-1} \in \{a, a^3\}$. If $bab^{-1} = a$, then a and b commute and G, whose elements are the $a^i b^k$, is abelian. Hence $bab^{-1} = a^{-1}$, and so $ba^j = a^{-j}b$ for all j. By Lemma 11.2.6, $b^2 \in N$, say $b^2 = a^j$; since b commutes with b^2, we get $a^j = ba^j b^{-1} = a^{-j}$, so $a^{2j} = e$ and $b^2 \in \{e, a^2\}$. Now

$$(a^i b^k)(a^j b^l) = a^{i+(-1)^k j}\, b^k b^l, \qquad b^k b^l \in \{e, b, b^2\},$$

so the multiplication of G is completely determined by whether $b^2 = e$ or $b^2 = a^2$.

- If $b^2 = e$, the same rules hold in D_4 for a the rotation by $\pi/2$ and b a reflection, whose products $a^i b^k$ are the eight elements of D_4. So $a^i b^k \mapsto a^i b^k$ is an isomorphism $G \cong D_4$.

- If $b^2 = a^2$, the same rules hold in Q_8 for $a = i$ and $b = j$: i has order 4, $j^2 = -1 = i^2$, and $jij^{-1} = ji(-j) = -(ji)j = kj = -i = i^{-1}$. The elements $i^m j^l$ are $1, i, -1, -i, j, k, -j, -k$, so $G \cong Q_8$. □

NOTE The class equation gives another way into this classification. If G is non-abelian of order 8, then $|Z(G)| \in \{2, 4\}$ by Theorem 11.2.4, and the argument of Corollary 11.2.5 rules out 4: for $g \notin Z(G)$, $Z(g)$ contains $Z(G)$ and g, so it would have more than 4 elements and be all of G. So $|Z(G)| = 2$. For each $g \notin Z(G)$, the centralizer $Z(g)$ lies strictly between $Z(G)$ and G, so it has order 4 and $|C_g| = 2$. The class equation is therefore $8 = 1+1+2+2+2$: two central elements and three classes of size 2. One can then work out the possible multiplications from there.

11.3 Finite Subgroups of $SO(3)$

Using group actions and counting, we now classify the finite groups of rotations of $\mathbb{R}^3$. Recall from Chapter 9 that $SO(3)$ is the group of orthogonal operators on $\mathbb{R}^3$ (with the standard inner product) of determinant 1.

11.3.1 Rotations and Symmetry Groups

Proposition 11.3.1 (Elements of $SO(3)$ are rotations)

Let $T \in SO(3)$ with $T \neq \mathrm{id}$. The vectors fixed by T form a line L, the ***axis*** of T, and T maps the plane $L^\perp$ to itself and acts on it as a rotation by an angle $\theta \in (0, 2\pi)$. In particular T fixes exactly two unit vectors $\pm v$, where $L = \mathbb{R}v$; we call them the ***poles*** of T.

Proof. By the structure theorem for orthogonal operators (Theorem 9.5.19), $\mathbb{R}^3$ is an orthogonal direct sum $V_1 \oplus V_{-1} \oplus W_1 \oplus \cdots \oplus W_k$, where $V_{\pm 1} = \ker(T \mp I)$ and T acts on each plane W_j by a rotation through an angle $\theta_j \in (0, \pi)$. A rotation of a plane has determinant 1, so $1 = \det T = (-1)^{\dim V_{-1}}$, and $\dim V_{-1}$ is even. Since $\dim V_1 + \dim V_{-1} + 2k = 3$ and $T \neq \mathrm{id}$, there are two possibilities: $\dim V_1 = 1$, $\dim V_{-1} = 2$, $k = 0$, where T acts on $V_1^\perp = V_{-1}$ as the rotation by π; or $\dim V_1 = 1$, $\dim V_{-1} = 0$, $k = 1$, where T acts on $V_1^\perp = W_1$ as a rotation by $\theta_1 \in (0, \pi)$. In both cases the fixed vectors of T are those of the line $L = V_1$. □

Given a subset $\Sigma \subseteq \mathbb{R}^3$, its ***rotation group*** $\{T \in SO(3) : T(\Sigma) = \Sigma\}$ is a subgroup of $SO(3)$. It may be infinite: if Σ is a circle centered at the origin, it contains every rotation about the axis perpendicular to the plane of the circle. For the following shapes it is finite.

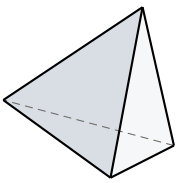
Example 11.3.2 (Cyclic and dihedral rotation groups)

Let Σ be a regular n-gon ($n \geqslant 3$) in the xy-plane, centered at the origin. Its rotation group consists of the n rotations about the z-axis by the angles $2\pi k/n$, and the n half-turns (rotations by π) about the lines in the xy-plane through the origin and a vertex or the midpoint of an edge. Each of these rotations maps the xy-plane to itself, and restricting to that plane is an isomorphism onto the symmetry group of the polygon: the rotations about the z-axis restrict to the rotations of the polygon and the half-turns to its reflections, and an element T is recovered from its restriction, since it acts on the z-axis by $\det T|_{xy\text{-plane}} = \pm 1$. So the rotation group of Σ is isomorphic to the dihedral group D_n (Definition 2.2.14).

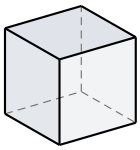
If instead Σ is a pyramid built on the n-gon, with apex on the positive z-axis, the half-turns are lost (they move the apex), and the rotation group is the cyclic group $\mathbb{Z}/n$ of rotations about the z-axis by the angles $2\pi k/n$. For every $n \geqslant 1$, these n rotations form a cyclic subgroup of $SO(3)$ of order n.

Example 11.3.3 (Regular polyhedra)

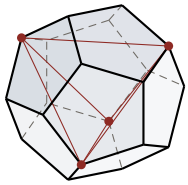
The rotation groups of the regular polyhedra centered at the origin are finite. The cube and the octahedron have the same rotation group: the centers of the faces of a cube are the vertices of an octahedron, and conversely, so a rotation preserving one preserves the other. In the same way the dodecahedron (12 pentagonal faces) and the icosahedron (20 triangular faces) have the same rotation group. The orders follow from the Orbit–Stabilizer Theorem applied to the faces: the rotation group acts transitively on the faces, and the stabilizer of a face with k sides consists of the k rotations about the axis through its center. This gives $4 \cdot 3 = 12$ for the tetrahedron, $6 \cdot 4 = 24$ for the cube and $12 \cdot 5 = 60$ for the dodecahedron. We call these three groups the ***tetrahedral***, ***octahedral*** and ***icosahedral*** groups.



tetrahedron
A_4, $(2,3,3)$



cube
S_4, $(2,3,4)$



dodecahedron
A_5, $(2,3,5)$

11.3.2 Poles and the Classification

Let $G \subseteq SO(3)$ be a finite subgroup, and let P be the set of poles of the elements of G other than e:

$$P = \{v \in \mathbb{R}^3 : \|v\| = 1 \text{ and } gv = v \text{ for some } g \in G,\ g \neq e\}.$$

If v is a pole of $g \neq e$ and $h \in G$, then $hgh^{-1} \neq e$ and $(hgh^{-1})(hv) = hgv = hv$; so the poles of hgh^{-1} are the images under h of the poles of g. In particular $hv \in P$, and G acts on the finite set P.

Lemma 11.3.4 (Stabilizers of poles)

Let $p \in P$. Then $\mathrm{Stab}(p)$ consists of the identity and the elements of G with axis $\mathbb{R}p$; in particular $\mathrm{Stab}(-p) = \mathrm{Stab}(p)$. Moreover $\mathrm{Stab}(p)$ is cyclic, of some order $r_p \geqslant 2$.

Proof. An element $g \neq e$ fixes p exactly when p is one of its poles, that is, when its axis is $\mathbb{R}p$; this condition is the same for $-p$. Each $g \in \mathrm{Stab}(p)$ maps the plane $p^\perp$ to itself (for an orthogonal operator, the orthogonal complement of an invariant subspace is invariant, by Proposition 9.5.4), and g is determined by its restriction to $p^\perp$, since $gp = p$. In an orthonormal basis of $\mathbb{R}^3$ whose first vector is p, the matrix of g is block diagonal with blocks 1 and the matrix of $g|_{p^\perp}$, so $\det g|_{p^\perp} = \det g = 1$. Restriction is therefore an injective homomorphism from $\mathrm{Stab}(p)$ to $SO(p^\perp) \cong SO(2)$, and $SO(2)$ is isomorphic to the circle group $\{z \in \mathbb{C} : |z| = 1\}$ (Proposition 9.5.17). Let $r = |\mathrm{Stab}(p)|$. By Lagrange's theorem every z in the image satisfies $z^r = 1$, and the polynomial $z^r - 1$ has at most r roots, namely the rth roots of unity $e^{2\pi i k/r}$. So the image is the group of rth roots of unity, which is cyclic, generated by $e^{2\pi i/r}$. Finally $r \geqslant 2$, because p is a pole of some $g \neq e$. □

For $p \in P$ we write $r_p = |\mathrm{Stab}(p)|$. By Proposition 11.1.1, poles in the same orbit have the same r_p.

Proposition 11.3.5 (The orbits of poles)

Let $G \neq \{e\}$ be a finite subgroup of $SO(3)$. The action of G on its poles has two or three orbits.

(1) If there are two orbits, each consists of a single pole, the two poles are $\pm p$, and $G = \mathrm{Stab}(p)$ is a cyclic group of rotations about the axis $\mathbb{R}p$.

(2) If there are three orbits $\mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3$, numbered so that the corresponding values $r_1 \leqslant r_2 \leqslant r_3$ of r_p are increasing, then (r_1, r_2, r_3) and $|G|$ are one

of

$$(2,2,n),\ |G| = 2n\ (n \geqslant 2); \qquad (2,3,3),\ |G| = 12;$$
$$(2,3,4),\ |G| = 24; \qquad (2,3,5),\ |G| = 60.$$

In the case $(2,2,n)$, $G \cong D_n$, where for $n = 2$ we write D_2 for the group $\mathbb{Z}/2 \times \mathbb{Z}/2$.

In the cases (1) and $(2,2,n)$, all elements of G of order greater than 2 are rotations about a single axis.

Proof. The set P is finite and nonempty, since each $g \neq e$ has two poles. Count the set

$$\Sigma = \{(g,p) \in G \times P : g \neq e,\ gp = p\}$$

in two ways. Each $g \neq e$ has exactly two poles, so $|\Sigma| = 2(|G|-1)$. Each $p \in P$ is fixed by the $r_p - 1$ elements of $\mathrm{Stab}(p) \setminus \{e\}$, so $|\Sigma| = \sum_{p \in P}(r_p - 1)$. Let $\mathcal{O}_1, \ldots, \mathcal{O}_k$ be the orbits and r_i the common value of r_p on $\mathcal{O}_i$; by the Orbit–Stabilizer Theorem $|\mathcal{O}_i| = |G|/r_i$. Hence

$$2(|G|-1) = \sum_{i=1}^{k} \frac{|G|}{r_i}(r_i - 1), \qquad \text{that is,} \qquad 2 - \frac{2}{|G|} = \sum_{i=1}^{k}\left(1 - \frac{1}{r_i}\right). \qquad (*)$$

Since $|G| \geqslant 2$, the left side lies in $[1,2)$, and since $r_i \geqslant 2$, each term on the right lies in $[\frac{1}{2}, 1)$. So $k/2 < 2$, that is, $k \leqslant 3$; and $k \neq 1$, since then the right side would be less than 1.

Two orbits. Here $(*)$ reads $\frac{1}{r_1} + \frac{1}{r_2} = \frac{2}{|G|}$. Each r_i is the order of a subgroup of G, so $r_i \leqslant |G|$ and $\frac{1}{r_i} \geqslant \frac{1}{|G|}$; equality in the sum forces $r_1 = r_2 = |G|$. Then each orbit has $|G|/r_i = 1$ element, so $P = \{p, p'\}$ with both poles fixed by all of G. Every $g \neq e$ has its two poles in P, so its poles are p and p', and $p' = -p$. So $G = \mathrm{Stab}(p)$, which is cyclic by Lemma 11.3.4, and consists of rotations about $\mathbb{R}p$.

Three orbits. Here $(*)$ reads

$$\frac{1}{r_1} + \frac{1}{r_2} + \frac{1}{r_3} = 1 + \frac{2}{|G|} > 1.$$

If $r_1 \geqslant 3$ the left side is at most $\frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1$, so $r_1 = 2$. If $r_2 \geqslant 4$ the left side is at most $\frac{1}{2} + \frac{1}{4} + \frac{1}{4} = 1$, so $r_2 \in \{2,3\}$.

Suppose $r_2 = 2$. Then $\frac{1}{r_3} = \frac{2}{|G|}$, so $n := r_3 = |G|/2$, and $n \geqslant r_2 = 2$. The orbit $\mathcal{O}_3$ has $|G|/n = 2$ elements, say q and q'. The group $\mathrm{Stab}(q)$ permutes $\mathcal{O}_3$ and fixes q, so it fixes q'; a non-identity element of $\mathrm{Stab}(q)$ has only the poles $\pm q$, so $q' = -q$. By Lemma 11.3.4, $\mathrm{Stab}(q) = \langle \rho \rangle$ is cyclic of order n, and it has index 2 in G. Let $g \in G \setminus \mathrm{Stab}(q)$. Since g preserves $\mathcal{O}_3 = \{q, -q\}$ and moves q, $gq = -q$, and then $g^2 q = q$. The poles $\pm v$ of g are not $\pm q$, because g moves

q and $-q$; they are fixed by g^2. If $g^2 \neq e$, its only poles would be $\pm q$; hence $g^2 = e$. So every element of $G \setminus \mathrm{Stab}(q)$ is a half-turn. Fix such an element t. Then $G = \mathrm{Stab}(q) \sqcup \mathrm{Stab}(q)t = \{\rho^k, \rho^k t : 0 \leqslant k < n\}$, with $\rho^n = t^2 = e$; and since $\rho t \notin \mathrm{Stab}(q)$, also $(\rho t)^2 = e$, that is, $t\rho t^{-1} = \rho^{-1}$. These relations determine the product of any two elements of G (using $t\rho^k = \rho^{-k}t$), and the rotation r and a reflection s of the dihedral group D_n satisfy the same relations; so $\rho^k t^l \mapsto r^k s^l$ is an isomorphism $G \cong D_n$. (For $n = 2$ the relations say that $G \cong \mathbb{Z}/2 \times \mathbb{Z}/2 = D_2$, the group of order 4 consisting of the identity and three half-turns.)

Suppose $r_2 = 3$. Then $\frac{1}{r_3} > 1 - \frac{1}{2} - \frac{1}{3} = \frac{1}{6}$, so $r_3 \in \{3, 4, 5\}$, and $\frac{2}{|G|} = \frac{1}{2} + \frac{1}{3} + \frac{1}{r_3} - 1$ gives $|G| = 12, 24, 60$ respectively.

Finally, in case (1) every element of G is a rotation about $\mathbb{R}p$, and in the case $(2, 2, n)$ every element outside $\mathrm{Stab}(q)$ has order 2. □

It remains to identify the groups in the last three cases. We use that a cycle of odd length k is a product of $k - 1$ transpositions, $(a_1\, a_2 \cdots a_k) = (a_1\, a_2)(a_2\, a_3) \cdots (a_{k-1}\, a_k)$, hence an even permutation.

Proposition 11.3.6 (The exceptional cases)

Let G be a finite subgroup of $SO(3)$ in one of the cases $(2, 3, 3)$, $(2, 3, 4)$, $(2, 3, 5)$ of Proposition 11.3.5. Then $G \cong A_4$, $G \cong S_4$, $G \cong A_5$, respectively.

Proof. Recall that $\mathrm{Stab}(-p) = \mathrm{Stab}(p)$, so p and $-p$ have the same r_p. When r_1, r_2, r_3 are distinct, each orbit $\mathcal{O}_i$ is therefore closed under $p \mapsto -p$, and consists of $|\mathcal{O}_i|/2$ axes.

Case $(2, 3, 3)$**.** Here $|G| = 12$ and $|\mathcal{O}_2| = 12/3 = 4$. An element $g \neq e$ fixes only two unit vectors, so the action of G on the four points of $\mathcal{O}_2$ is faithful, and G is isomorphic to a subgroup H of S_4 of order 12. Since H has index 2, it contains the eight 3-cycles of S_4 by Lemma 11.2.6. These are even, so $H \cap A_4$ is a subgroup with at least 9 elements whose order divides 12; hence $H \cap A_4 = H = A_4$.

Case $(2, 3, 4)$**.** Here $|G| = 24$, and $\mathcal{O}_2$ consists of 8 poles forming a set X of four axes. The group G acts on X, since g maps the line $\mathbb{R}p$ to $\mathbb{R}(gp)$; let $\varphi : G \to \mathrm{Sym}(X) \cong S_4$ be this action. Let $p \in \mathcal{O}_2$ and let ρ generate $\mathrm{Stab}(p) \cong \mathbb{Z}/3$. The only line that ρ maps to itself is $\mathbb{R}p$: if $\rho w = \pm w$ for a unit vector w, then w is fixed by ρ or by ρ^2, both of which are non-identity elements with axis $\mathbb{R}p$, so $w = \pm p$. Hence ρ fixes the axis $\mathbb{R}p \in X$ and permutes the other three axes without fixed points; since $\rho^3 = e$, it permutes them in a 3-cycle. So for each axis $\ell \in X$, the image $\varphi(G)$ contains a 3-cycle fixing ℓ and its inverse, and $\varphi(G)$ contains all eight 3-cycles of S_4. Thus $|\varphi(G)| \geqslant 9$, so $|\varphi(G)| \in \{12, 24\}$ and $|\ker \varphi| \leqslant 2$. Suppose $\ker \varphi = \{e, h\}$. As $\ker \varphi$ is normal, $ghg^{-1} = h$ for every $g \in G$, so each g maps the axis of h to itself. Taking $g = \rho$ as above shows that the axis of h is $\mathbb{R}p$, for

every $p \in \mathcal{O}_2$; this is impossible, since $\mathcal{O}_2$ contains four different axes. So φ is injective, and $G \cong \varphi(G) = S_4$.

Case $(2,3,5)$**.** Here $|G| = 60$, and $\mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3$ consist of $30, 20, 12$ poles, forming $15, 10, 6$ axes. Every $g \neq e$ has exactly one axis $\mathbb{R}p$, and the elements with that axis are the $r_p - 1$ non-identity elements of $\mathrm{Stab}(p) \cong \mathbb{Z}/r_p$. Hence G consists of the identity, 15 half-turns (one about each axis of $\mathcal{O}_1$), 20 elements of order 3 (two about each axis of $\mathcal{O}_2$) and 24 elements of order 5 (four about each axis of $\mathcal{O}_3$); in particular G has no element of order 4.

For $q \in \mathcal{O}_1$, let h_q be the half-turn with axis $\mathbb{R}q$, so that $\mathrm{Stab}(q) = \{e, h_q\}$. For $g \in G$, gh_qg^{-1} is a non-identity element fixing gq, so $gh_qg^{-1} = h_{gq}$. Since $\mathcal{O}_1$ is an orbit, the 15 half-turns form one conjugacy class, and the centralizer $V_q = Z(h_q)$ has order $60/15 = 4$. Its elements other than e have order 2 (there are no elements of order 4), so V_q consists of e, h_q and two other half-turns, and it is abelian, because in a group in which $x^2 = e$ for all x we have $xy = (xy)^{-1} = y^{-1}x^{-1} = yx$. If h' is a half-turn in V_q, then every element of V_q commutes with h', so $V_q \subseteq Z(h')$, and the two groups are equal since both have order 4. So the 15 half-turns fall into five disjoint triples, and we obtain five subgroups $V_1, \ldots, V_5$, each of which is the centralizer of each of its three half-turns.

Since x commutes with h if and only if gxg^{-1} commutes with ghg^{-1}, we have $g\,Z(h)\,g^{-1} = Z(ghg^{-1})$. So G acts on $\{V_1, \ldots, V_5\}$ by conjugation, transitively because all half-turns are conjugate. Let K be the kernel of this action; it is a normal subgroup of G contained in the stabilizer of V_1, which has order $60/5 = 12$, so $|K|$ divides 12. If K contained a half-turn, it would contain all 15 half-turns, since K is normal; this is too many. By Lagrange's theorem K contains no element of order 5. If K contained an element ρ of order 3, with axis $\mathbb{R}p$ for some $p \in \mathcal{O}_2$, then for every $g \in G$ it would contain $g\rho g^{-1}$, a non-identity element of $\mathrm{Stab}(gp) \cong \mathbb{Z}/3$, and hence both elements of order 3 about the axis $\mathbb{R}(gp)$. As gp runs over $\mathcal{O}_2$, this gives all 20 elements of order 3, again too many. So $K = \{e\}$, and G is isomorphic to a subgroup H of S_5 of order 60. Since H has index 2, it contains the twenty 3-cycles and the twenty-four 5-cycles of S_5 by Lemma 11.2.6. These are even, so $H \cap A_5$ is a subgroup with at least 45 elements whose order divides 60; hence $H = A_5$. □

Theorem 11.3.7 (Finite subgroups of SO(3))

Every finite subgroup of $SO(3)$ is isomorphic to one of the following: a cyclic group $\mathbb{Z}/n$ ($n \geqslant 1$), a dihedral group D_n ($n \geqslant 2$), A_4, S_4, or A_5. Each of them occurs; in particular A_4, S_4 and A_5 are the tetrahedral, octahedral and icosahedral groups.

Proof. The first statement follows from Propositions 11.3.5 and 11.3.6. The groups $\mathbb{Z}/n$ and D_n for $n \geqslant 3$ occur by Example 11.3.2, and D_2 occurs as the group of the identity and the half-turns about the three coordinate axes. The tetrahedral group has order 12 (Example 11.3.3) and contains rotations of order 3 about four different axes, those through the vertices. By the last statement of Proposition 11.3.5, it is neither of type (1) nor of type $(2,2,n)$; since its order is 12, it is of type $(2,3,3)$, and it is isomorphic to A_4. In the same way the octahedral group, of order 24, contains rotations of order 4 about the three axes through the centers of opposite faces of the cube, so it is of type $(2,3,4)$ and isomorphic to S_4; and the icosahedral group, of order 60, contains rotations of order 5 about the six axes through the centers of opposite faces of the dodecahedron, so it is of type $(2,3,5)$ and isomorphic to A_5. □

NOTE For a regular polyhedron the poles are the vertices, the centers of the faces and the midpoints of the edges (scaled to unit length). Poles at midpoints of edges have $r = 2$; poles at vertices have r equal to the number of faces meeting at a vertex; and poles at centers of faces have r equal to the number of edges of a face. This gives $(2,3,3)$ for the tetrahedron (whose vertices and face centers form two different orbits with $r = 3$), $(2,3,4)$ for the cube and the octahedron, and $(2,3,5)$ for the dodecahedron and the icosahedron, as in the figure.

The proof of Proposition 11.3.6 becomes concrete for the cube and the dodecahedron.

Example 11.3.8 (The cube and its four diagonals)

Place the cube with vertices $(\pm1,\pm1,\pm1)$. Its vertices are the poles with $r = 3$, so in the proof of Proposition 11.3.6 the four axes in X are the four long diagonals of the cube, and the rotation group of the cube is S_4 acting on its diagonals. One can also see directly that only the identity maps each diagonal to itself. The diagonals are spanned by $v_1 = (1,1,1)$, $v_2 = (1,1,-1)$, $v_3 = (1,-1,1)$ and $v_4 = (-1,1,1)$. If a rotation T maps each line $\mathbb{R}v_i$ to itself, then $Tv_i = \varepsilon_i v_i$ with $\varepsilon_i = \pm1$, because T preserves lengths. Since $v_1 = v_2 + v_3 + v_4$,

$$\varepsilon_1(v_2 + v_3 + v_4) = Tv_1 = \varepsilon_2 v_2 + \varepsilon_3 v_3 + \varepsilon_4 v_4,$$

and as v_2, v_3, v_4 are linearly independent, $\varepsilon_1 = \varepsilon_2 = \varepsilon_3 = \varepsilon_4$. So $T = \pm\mathrm{id}$, and $\det T = 1$ forces $T = \mathrm{id}$.

Example 11.3.9 (The icosahedral group and five tetrahedra)

For the dodecahedron, the poles with $r = 5$ are the centers of the 12 faces, those with $r = 3$ are the 20 vertices, and those with $r = 2$ are the midpoints of the 30 edges. The five subgroups $V_1, \dots, V_5$ of the proof of Proposition 11.3.6 have a geometric description, which the proof does not need. The 20 vertices of the dodecahedron split into five sets of four, each the vertex set of a regular tetrahedron inscribed in the dodecahedron (one of them is marked in the figure), and the rotations of the dodecahedron permute these five tetrahedra. Each V_i consists of the identity and the three half-turns about the axes through the midpoints of opposite edges of one of the five tetrahedra. Since an isomorphism preserves the orders of elements, and the elements of A_5 of orders 5, 3 and 2 are the 5-cycles, the 3-cycles and the products $(ij)(k\ell)$ of two disjoint transpositions, the isomorphism with A_5 matches

- the 24 rotations about centers of faces with the 24 five-cycles,
- the 20 rotations about vertices with the 20 three-cycles,
- the 15 half-turns about midpoints of edges with the 15 elements $(ij)(k\ell)$.

11.4 Conjugacy in S_n and A_n

11.4.1 Conjugacy Classes in S_n

By Theorem 3.4.6, two permutations are conjugate in S_n if and only if they have the same cycle type; the key computation is

$$\tau\,(a_1\,a_2\,\cdots\,a_k)\,\tau^{-1} = (\tau(a_1)\,\tau(a_2)\,\cdots\,\tau(a_k)).$$

So the conjugacy classes of S_n correspond to the partitions of n, and the class equation of S_n is obtained by counting the permutations of each cycle type.

Example 11.4.1 (Conjugacy classes in S_3 and S_4)

For $n = 3$, the partitions $1+1+1$, $2+1$ and 3 give the identity, the 3 transpositions and the 2 three-cycles: $6 = 1+3+2$.
For $n = 4$, the five partitions give the identity (1 element), the transpositions ($\binom{4}{2} = 6$), the products $(ij)(k\ell)$ of two disjoint transpositions (3), the 3-cycles ($4 \cdot 2 = 8$) and the 4-cycles ($3! = 6$). The class equation of S_4 is $24 = 1+6+3+8+6$.

Example 11.4.2 (Normal subgroups of S_4)

A normal subgroup of S_4 is a union of conjugacy classes containing $\{e\}$, and its order divides 24. Adding 1 to sums of distinct class sizes among $6, 3, 8, 6$, the only divisors of 24 we obtain besides 1 and 24 are $1+3=4$ and $1+3+8=12$. Both unions are subgroups. The first is $V_4 = \{e, (1\,2)(3\,4), (1\,3)(2\,4), (1\,4)(2\,3)\}$, the ***Klein four-group***: each non-identity element has order 2, and the product of two distinct ones is the third (for instance $(1\,2)(3\,4)\cdot(1\,3)(2\,4) = (1\,4)(2\,3)$), so $V_4 \cong \mathbb{Z}/2 \times \mathbb{Z}/2$. The second consists of the identity, the elements $(i\,j)(k\,\ell)$ and the 3-cycles, which are exactly the even permutations; it is A_4. So the normal subgroups of S_4 are $\{e\}$, V_4, A_4 and S_4.

11.4.2 Conjugacy Classes in A_n

Recall the sign homomorphism $\operatorname{sgn} : S_n \to \{\pm 1\}$ (Definition 3.4.10) and the alternating group $A_n = \ker(\operatorname{sgn})$ (Definition 3.4.15), a normal subgroup of index 2 in S_n for $n \geqslant 2$. Transpositions are odd, and a k-cycle is a product of $k-1$ transpositions, $(i_1\, i_2 \cdots i_k) = (i_1\, i_2)(i_2\, i_3)\cdots(i_{k-1}\, i_k)$, so it has sign $(-1)^{k-1}$. Hence $\sigma \in A_n$ if and only if its cycle decomposition has an even number of cycles of even length.

A conjugacy class of S_n contained in A_n is a union of conjugacy classes of A_n, but it need not be a single class: conjugating by elements of A_n only, we may reach fewer elements.

Proposition 11.4.3 (Splitting of conjugacy classes in A_n)

Let C be a conjugacy class of S_n, $n \geqslant 2$. Then either $C \cap A_n = \emptyset$ or $C \subseteq A_n$. In the second case, let $\sigma \in C$. Then exactly one of the following holds.

(a) Some odd permutation commutes with σ, that is, $Z_{S_n}(\sigma) \not\subseteq A_n$; then C is a single conjugacy class of A_n.

(b) $Z_{S_n}(\sigma) \subseteq A_n$; then C splits into two conjugacy classes of A_n, each with $|C|/2$ elements.

Proof. Since A_n is normal, it contains the S_n-conjugacy class of each of its elements; so if C meets A_n, then $C \subseteq A_n$. Suppose $C \subseteq A_n$, and fix an odd permutation τ. Then $S_n = A_n \sqcup A_n\tau$, so

$$C = \{h\sigma h^{-1} : h \in A_n\} \cup \{(h\tau)\sigma(h\tau)^{-1} : h \in A_n\} = C' \cup C'',$$

where C' is the A_n-conjugacy class of σ and C'' is the A_n-conjugacy class of $\tau\sigma\tau^{-1}$. Two conjugacy classes of A_n are equal or disjoint, and $C' = C''$ if and only

if $\sigma \in C''$, that is, $\sigma = g\sigma g^{-1}$ for some $g = h\tau$ with $h \in A_n$. The elements $h\tau$ are exactly the odd permutations, so $C' = C''$ if and only if some odd permutation commutes with σ. In case (b), conjugation by τ maps C' into C'' (it sends $h\sigma h^{-1}$ to $(\tau h\tau^{-1})(\tau\sigma\tau^{-1})(\tau h\tau^{-1})^{-1}$, and $\tau h\tau^{-1} \in A_n$), and conjugation by τ^{-1} maps C'' into C' in the same way. These maps are inverse to each other, so $|C'| = |C''| = |C|/2$. □

Proposition 11.4.4 (When a class splits)

Let $\sigma \in A_n$, and count the fixed points of σ as cycles of length 1. The S_n-conjugacy class of σ splits into two conjugacy classes of A_n if and only if the cycle lengths of σ are all odd and pairwise distinct. (In particular, if it splits, σ has at most one fixed point.)

Proof. Suppose first that σ has a cycle c of even length. Then c commutes with σ (it commutes with itself and with the other cycles of σ, which are disjoint from it), and c is odd. Suppose next that σ has two cycles $(a_1 \cdots a_k)$ and $(b_1 \cdots b_k)$ of the same odd length k; for $k = 1$ these are two fixed points a_1 and b_1. Then $\tau = (a_1\, b_1)(a_2\, b_2)\cdots(a_k\, b_k)$ is a product of k disjoint transpositions, hence odd, and by the key computation $\tau\sigma\tau^{-1} = \sigma$: conjugation by τ exchanges the two cycles and fixes the other cycles of σ, whose entries τ does not move. In both cases $Z_{S_n}(\sigma) \not\subseteq A_n$, and the class does not split by Proposition 11.4.3.

Conversely, suppose the cycle lengths of σ are odd and pairwise distinct, let $X_1, \ldots, X_\ell$ be the supports of the cycles $c_1, \ldots, c_\ell$ of σ (including the 1-cycles), and let τ commute with σ. By the key computation, the cycles of $\tau\sigma\tau^{-1} = \sigma$ are obtained by applying τ to the entries of the cycles of σ; by the uniqueness of the cycle decomposition (Theorem 3.4.1), τ maps each X_j onto the support of a cycle of σ of the same length, which is X_j itself because the lengths are distinct. Write $c_j = (a_1\, a_2 \cdots a_k)$. Then $(\tau(a_1)\,\tau(a_2)\cdots\tau(a_k)) = c_j$, so $\tau(a_{i+1}) = c_j(\tau(a_i))$ for each i; if $\tau(a_1) = c_j^m(a_1)$, it follows that $\tau(a_i) = c_j^{i-1}(\tau(a_1)) = c_j^m(a_i)$ for all i. So τ agrees with $c_j^{m_j}$ on X_j for some m_j, and $\tau = c_1^{m_1}\cdots c_\ell^{m_\ell}$. Each c_j is a cycle of odd length, hence even, so τ is even. Thus $Z_{S_n}(\sigma) \subseteq A_n$, and the class splits by Proposition 11.4.3. □

Example 11.4.5 (Conjugacy classes of A_4 and A_5)

In A_4, the three elements $(i\,j)(k\,\ell)$ form one class (they have cycles of even length), while the eight 3-cycles, with cycle lengths 3, 1, split into two classes of 4; for instance $(1\,2\,3)$ and $(1\,3\,2) = (2\,3)(1\,2\,3)(2\,3)^{-1}$ are not conjugate in A_4. The class equation of A_4 is $12 = 1 + 3 + 4 + 4$. The Klein four-group V_4 is normal in S_4, hence in A_4; so A_4 is not simple.

In A_5, the 15 elements $(i\,j)(k\,\ell)$ form one class (for instance, the odd permutation $(1\,2)$ commutes with $(1\,2)(3\,4)$), and so do the 20 three-cycles, with cycle lengths $3, 1, 1$ (for instance, $(4\,5)$ commutes with $(1\,2\,3)$). The 24 five-cycles split into two classes of 12. The class equation of A_5 is

$$60 = 1 + 15 + 20 + 12 + 12.$$

Corollary 11.4.6 (3-cycles are conjugate in A_n)

For $n \geqslant 5$, the 3-cycles form a single conjugacy class of A_n.

Proof. The 3-cycles form a conjugacy class of S_n, contained in A_n. A 3-cycle has $n - 3 \geqslant 2$ fixed points, that is, two cycles of length 1, so by Proposition 11.4.4 this class does not split. □

For $n = 4$ the corollary fails, as the example shows; this is where the hypothesis $n \geqslant 5$ of the next theorem enters.

11.4.3 Simplicity of A_n

Recall that a group is *simple* if its only normal subgroups are $\{e\}$ and the group itself.

Theorem 11.4.7 (Simplicity of A_5)

The alternating group A_5 is simple.

Proof. A normal subgroup H of A_5 is a union of conjugacy classes of A_5 containing $\{e\}$, and $|H|$ divides 60. By Example 11.4.5, $|H| = 1 + s$, where s is a sum of distinct terms of the list $15, 20, 12, 12$. The possible values of $1 + s$ are $1, 13, 16, 21, 25, 28, 33, 36, 40, 45, 48, 60$, and of these only 1 and 60 divide 60. So $H = \{e\}$ or $H = A_5$. □

For larger n we argue with 3-cycles.

Lemma 11.4.8 (A_n is generated by 3-cycles)

For $n \geqslant 3$, every element of A_n is a product of 3-cycles.

Proof. By induction on n. For $n = 3$, $A_3 = \{e, (1\,2\,3), (1\,3\,2)\}$ and the claim is clear. Let $n \geqslant 4$, assume the claim for A_{n-1}, and let $\sigma \in A_n$. If $\sigma(n) = n$, then σ is an even permutation of $\{1, \ldots, n-1\}$, so it is a product of 3-cycles by induction. Otherwise, let $i = \sigma(n) \neq n$ and let j be any element distinct from i and n. The

3-cycle $(j\,i\,n)$ sends i to n, so $\tau = (j\,i\,n)\sigma \in A_n$ satisfies $\tau(n) = n$, and τ is a product of 3-cycles by the first case. Hence so is $\sigma = (j\,i\,n)^{-1}\tau = (n\,i\,j)\tau$. □

Theorem 11.4.9 (Simplicity of A_n)

For every $n \geqslant 5$, the alternating group A_n is simple.

Proof. Let H be a normal subgroup of A_n with $H \neq \{e\}$. If H contains a 3-cycle, it contains all 3-cycles, since they form one conjugacy class of A_n (Corollary 11.4.6); then $H = A_n$ by Lemma 11.4.8. So it suffices to find a 3-cycle in H. Throughout, we use that for $\sigma \in H$ and $g \in A_n$, the element $g\sigma g^{-1}\sigma^{-1}$ lies in H, because $g\sigma g^{-1} \in H$. Products are composed from right to left.

Choose $\sigma \in H$, $\sigma \neq e$. If σ has order m and p is a prime dividing m, then $\sigma^{m/p} \in H$ has order p; so we may assume that σ has prime order p. Its disjoint cycles $c_1, \ldots, c_\ell$ commute, so $e = \sigma^p = c_1^p \cdots c_\ell^p$; the c_i^p move disjoint sets of points, so each $c_i^p = e$. A k-cycle has order k, so every cycle of σ of length greater than 1 has length p: σ is a product of disjoint p-cycles.

Case 1: $p \geqslant 5$. Write $\sigma = (i_1\,i_2\,i_3\,i_4\,i_5 \cdots i_p)\tau$, where τ is the product of the other p-cycles of σ and fixes $i_1, \ldots, i_p$. Let $g = (i_4\,i_3\,i_2)$, so $g^{-1} = (i_2\,i_3\,i_4)$. Applying σ^{-1}, g^{-1}, σ, g in turn,

$$\begin{aligned} i_1 &\mapsto i_p \mapsto i_p \mapsto i_1 \mapsto i_1, \\ i_2 &\mapsto i_1 \mapsto i_1 \mapsto i_2 \mapsto i_4, \\ i_3 &\mapsto i_2 \mapsto i_3 \mapsto i_4 \mapsto i_3, \\ i_4 &\mapsto i_3 \mapsto i_4 \mapsto i_5 \mapsto i_5, \\ i_5 &\mapsto i_4 \mapsto i_2 \mapsto i_3 \mapsto i_2. \end{aligned}$$

Any other point x is fixed: neither x nor $\sigma^{-1}(x)$ lies in $\{i_2, i_3, i_4\}$, so g^{-1} and g do not move them. Hence

$$g\sigma g^{-1}\sigma^{-1} = (i_2\,i_4\,i_5) \in H.$$

Case 2: $p = 3$. If σ is a 3-cycle we are done. Otherwise $\sigma = (i_1\,i_2\,i_3)(i_4\,i_5\,i_6)\tau$, with τ fixing $i_1, \ldots, i_6$. With $g = (i_4\,i_3\,i_2)$, the same kind of computation gives $g\sigma g^{-1}\sigma^{-1} = (i_1\,i_5\,i_2\,i_4\,i_3)$, an element of H of order 5, and Case 1 applies to it.

Case 3: $p = 2$, and σ is a product of two transpositions, $\sigma = (i_1\,i_2)(i_3\,i_4)$. Since $n \geqslant 5$, there is a point $i_5 \notin \{i_1, i_2, i_3, i_4\}$. With $g = (i_5\,i_3\,i_1)$, we get $g\sigma g^{-1}\sigma^{-1} = (i_1\,i_5\,i_2\,i_4\,i_3)$, of order 5, and Case 1 applies.

Case 4: $p = 2$, and σ is a product of more than two transpositions. Since σ is even, it has at least four, and $\sigma = (i_1\,i_2)(i_3\,i_4)(i_5\,i_6)\tau$ with τ fixing $i_1, \ldots, i_6$. With $g = (i_5\,i_3\,i_1)$, we get $g\sigma g^{-1}\sigma^{-1} = (i_1\,i_5\,i_3)(i_2\,i_4\,i_6)$, an element of H of order 3, and Case 2 applies. □

Corollary 11.4.10 (Normal subgroups of S_n)
For $n \geqslant 5$, the only normal subgroups of S_n are $\{e\}$, A_n and S_n.

Proof. Let $H \trianglelefteq S_n$. Then $H \cap A_n$ is normal in S_n, hence in A_n, so $H \cap A_n$ is $\{e\}$ or A_n by Theorem 11.4.9. If $A_n \subseteq H$, then $H = A_n$ or $H = S_n$, since A_n has index 2. If $H \cap A_n = \{e\}$, then H contains at most one odd permutation (if $h_1, h_2 \in H$ are odd, then $h_1^{-1}h_2 \in H \cap A_n$, so $h_1 = h_2$), and $|H| \leqslant 2$. Suppose $H = \{e, \sigma\}$ with $\sigma \neq e$. Since H is normal, $\tau\sigma\tau^{-1} = \sigma$ for every $\tau \in S_n$. But σ has a cycle $(a_1\ a_2\ \cdots)$ of length at least 2, and as $n \geqslant 3$ there is a point $b \notin \{a_1, a_2\}$; for $\tau = (a_2\ b)$, $\tau\sigma\tau^{-1}(a_1) = \tau(\sigma(a_1)) = \tau(a_2) = b \neq \sigma(a_1)$, a contradiction. So $H = \{e\}$. □

11.5 The Sylow Theorems

11.5.1 The Statements

By Lagrange's theorem, the order of a subgroup of G divides $|G|$. The converse fails: a divisor of $|G|$ need not be the order of any subgroup.

Example 11.5.1 (Missing subgroups)
A subgroup of index 2 contains every element of odd order (Lemma 11.2.6). So a subgroup of order 6 in A_4 would contain the identity and the eight 3-cycles, and a subgroup of order 30 in A_5 would contain the identity, the twenty 3-cycles and the twenty-four 5-cycles. Both are impossible: A_4 has no subgroup of order 6, and A_5 has no subgroup of order 30.

The Sylow theorems show that the converse does hold for prime powers, and they say much more.

Definition 11.5.2 (Sylow p-subgroup)
Let G be a finite group and p a prime, and write $|G| = p^e m$ with $p \nmid m$. A ***Sylow*** p***-subgroup*** of G is a subgroup of order p^e. A p***-subgroup*** of G is a subgroup that is a p-group.

Theorem 11.5.3 (The Sylow Theorems (Sylow, 1872))
Let G be a finite group and p a prime, and write $|G| = p^e m$ with $p \nmid m$.

(1) G has a Sylow p-subgroup.

(2) Any two Sylow p-subgroups H and H' of G are conjugate: $H' = gHg^{-1}$ for some $g \in G$. More generally, every p-subgroup of G is contained in a Sylow p-subgroup.
(3) The number s_p of Sylow p-subgroups of G satisfies $s_p \equiv 1 \pmod{p}$ and $s_p \mid m$.

We prove these theorems in Section 11.5.3. When $p \nmid |G|$ they are trivial ($e = 0$, and $\{e\}$ is the only Sylow p-subgroup), so the interesting case is $p \mid |G|$. A few remarks:

- Since $s_p \equiv 1 \pmod{p}$, s_p is prime to p, so $s_p \mid m$ is equivalent to $s_p \mid |G|$.
- A conjugate of a Sylow p-subgroup is again a Sylow p-subgroup. So by (2), a Sylow p-subgroup H is normal if and only if it is the only one, that is, $s_p = 1$: if $s_p = 1$ then $gHg^{-1} = H$ for all g, and if H is normal then every Sylow p-subgroup, being conjugate to H, equals H.

Corollary 11.5.4 (Cauchy's theorem)

If p is a prime dividing $|G|$, then G contains an element of order p.

Proof. By Sylow's first theorem, G has a subgroup H of order $p^e \geqslant p$. Pick $g \in H$, $g \neq e$. The order of g divides $|H| = p^e$, so it is p^k for some $1 \leqslant k \leqslant e$. Then $g^{p^{k-1}}$ has order p. □

NOTE For a p-group G the Sylow theorems say nothing: the only Sylow p-subgroup is G itself. In the Sylow approach to classifying finite groups the p-groups are therefore the hardest case, and their number grows very fast with the exponent. For $p = 2$:

order	2	4	8	16	32
number of groups	1	2	5	14	51

(order 4 is Corollary 11.2.5 and order 8 is Proposition 11.2.7), and there are already 56092 groups of order $2^8 = 256$.

11.5.2 Normalizers and Conjugate Subgroups

A group G acts by conjugation on the set of its subgroups, $g \cdot H = gHg^{-1}$. The orbit of H is the set of subgroups conjugate to H, and H is normal exactly when this orbit is $\{H\}$.

Definition 11.5.5 (Normalizer)

The ***normalizer*** of a subgroup $H \leqslant G$ is

$$N_G(H) = \{g \in G : gHg^{-1} = H\},$$

the stabilizer of H for the conjugation action on subgroups. We write $N(H)$ when G is clear.

As a stabilizer, $N_G(H)$ is a subgroup of G, and it contains H. For a subgroup K with $H \leqslant K \leqslant G$, H is normal in K if and only if $kHk^{-1} = H$ for all $k \in K$, that is, if and only if $K \subseteq N_G(H)$. So $N_G(H)$ is the largest subgroup of G in which H is normal, and $H \trianglelefteq G$ if and only if $N_G(H) = G$. The normalizer measures how far H is from being normal.

Example 11.5.6 (Normalizers in S_3)

Let $\sigma = (1\,2\,3)$ and $\tau = (1\,2)$. The subgroup $A_3 = \{e, \sigma, \sigma^2\}$ is normal, so $N(A_3) = S_3$, even though τ does not commute with σ: $\tau\sigma\tau^{-1} = (2\,1\,3) = \sigma^2$. Conjugation by τ moves the elements of A_3 but preserves the subgroup. On the other hand, for $H = \{e, \tau\}$ we have $gHg^{-1} = \{e, g\tau g^{-1}\}$, which equals H if and only if g commutes with τ; the centralizer of τ has $6/3 = 2$ elements, so $N(H) = \{e, \tau\} = H$.

Proposition 11.5.7 (Counting conjugate subgroups)

The number of subgroups of G conjugate to H is $[G : N_G(H)]$.

Proof. These subgroups form the orbit of H for the conjugation action, whose stabilizer is $N_G(H)$. Apply the Orbit–Stabilizer Theorem. □

11.5.3 Proofs of the Sylow Theorems

Fix a finite group G and a prime p, and write $|G| = n = p^e m$ with $p \nmid m$. The first two theorems are proved by letting G act on suitable sets by left multiplication, and the third by letting a Sylow subgroup act by conjugation.

Lemma 11.5.8 (A binomial coefficient)

If $n = p^e m$ with $p \nmid m$, then $p \nmid \binom{n}{p^e}$.

Proof. We have

$$\binom{n}{p^e} = \frac{n(n-1)\cdots(n-p^e+1)}{p^e(p^e-1)\cdots 1} = \prod_{k=0}^{p^e-1} \frac{p^e m - k}{p^e - k}.$$

The factor for $k = 0$ is m, which is prime to p. For $0 < k < p^e$, write $k = p^a k'$ with $p \nmid k'$; then $a < e$, and

$$p^e m - k = p^a(p^{e-a}m - k'), \qquad p^e - k = p^a(p^{e-a} - k'),$$

where neither $p^{e-a}m - k'$ nor $p^{e-a} - k'$ is divisible by p, since both are congruent to $-k'$ modulo p. So the numerator and the denominator of each factor are divisible by exactly the same power of p. Hence the power of p in the numerator of the product equals the power of p in its denominator, and the integer $\binom{n}{p^e}$ is not divisible by p. □

Lemma 11.5.9 (Stabilizers of subsets)

Let G act on the set of its subsets by left multiplication, $g \cdot U = gU$. For every subset $U \subseteq G$, $|\mathrm{Stab}(U)|$ divides $|U|$.

Proof. Let $H = \mathrm{Stab}(U) = \{g \in G : gU = U\}$. Then $hu \in U$ for all $h \in H$ and $u \in U$, so H acts on U by left multiplication. The orbit of $u \in U$ is the right coset Hu, which has $|H|$ elements. Since U is the disjoint union of these orbits, $|H|$ divides $|U|$. □

Proof of the First Sylow Theorem. Let S be the set of all subsets of G with p^e elements, so $|S| = \binom{n}{p^e}$. Left multiplication preserves the number of elements of a subset, so G acts on S by $g \cdot U = gU$. By Lemma 11.5.8, $p \nmid |S|$. Since $|S|$ is the sum of the sizes of the orbits, some orbit $G \cdot U$ has $p \nmid |G \cdot U|$. Now $|G| = |G \cdot U| \cdot |\mathrm{Stab}(U)|$ is divisible by p^e, and $p \nmid |G \cdot U|$, so p^e divides $|\mathrm{Stab}(U)|$. On the other hand $|\mathrm{Stab}(U)|$ divides $|U| = p^e$ by Lemma 11.5.9. So $|\mathrm{Stab}(U)| = p^e$, and $\mathrm{Stab}(U)$ is a Sylow p-subgroup. (In fact U is then a single orbit of $\mathrm{Stab}(U)$, that is, a right coset of $\mathrm{Stab}(U)$.) □

Proof of the Second Sylow Theorem. Let H be a Sylow p-subgroup, which exists by the first theorem, and let K be a p-subgroup of G. Let G act on the set G/H of left cosets of H by left multiplication; $|G/H| = m$ is not divisible by p. Restrict this action to K. The size of each K-orbit divides $|K|$, so it is a power of p. Since $p \nmid m$ and m is the sum of the sizes of the K-orbits, some orbit has size $p^0 = 1$: there is a coset gH with $kgH = gH$ for all $k \in K$. Then $g^{-1}kg \in H$, that is, $k \in gHg^{-1}$, for every $k \in K$. So $K \subseteq gHg^{-1}$, which is a Sylow p-subgroup, since it has the same order as H. If K is itself a Sylow p-subgroup, then $|K| = |gHg^{-1}|$ and $K = gHg^{-1}$. □

Proof of the Third Sylow Theorem. Let G act by conjugation on the set of its Sylow p-subgroups. By the second theorem this action is transitive, and the stabilizer of a Sylow p-subgroup H is $N(H)$, so $s_p = |G|/|N(H)|$ (Proposition 11.5.7). Since $H \leqslant N(H)$, $|N(H)| = p^e t$ for some integer t, and $s_p = p^e m/(p^e t) = m/t$ divides m.

Now restrict the conjugation action to H. The orbit of H itself is $\{H\}$. We claim that it is the only orbit of size 1. Suppose H' is a Sylow p-subgroup with $hH'h^{-1} = H'$ for all $h \in H$, that is, $H \subseteq N(H')$. The order of $N(H')$ is a multiple of $|H'| = p^e$ and divides $|G| = p^e m$, so p^e is the largest power of p dividing $|N(H')|$, and H and H' are Sylow p-subgroups of $N(H')$. By the second theorem applied to the group $N(H')$, $H = xH'x^{-1}$ for some $x \in N(H')$. But $xH'x^{-1} = H'$ by the definition of $N(H')$, so $H = H'$. This proves the claim.

Every orbit of the H-action has size dividing $|H| = p^e$, so every orbit other than $\{H\}$ has size divisible by p. Since s_p is the sum of the sizes of the orbits, $s_p \equiv 1 \pmod{p}$. □

11.6 Semidirect Products

11.6.1 Semidirect and Direct Products

If N is a normal subgroup of G, we have an exact sequence $1 \to N \to G \to G/N \to 1$ (Definition 3.5.1). This does not make G isomorphic to $N \times G/N$, and G need not even contain a subgroup isomorphic to G/N: in $0 \to p\mathbb{Z} \to \mathbb{Z} \to \mathbb{Z}/p \to 0$, the group $\mathbb{Z}$ has no nonzero element of finite order, so no subgroup isomorphic to $\mathbb{Z}/p$.

Suppose now that $H \leqslant G$ is a subgroup that contains exactly one element of each coset of N. Then every $g \in G$ can be written uniquely as $g = nh$ with $n \in N$ and $h \in H$, so $(n, h) \mapsto nh$ is a bijection $N \times H \to G$. It need not be a homomorphism, but since N is normal,

$$(n_1 h_1)(n_2 h_2) = \bigl(n_1 \cdot h_1 n_2 h_1^{-1}\bigr)(h_1 h_2), \qquad h_1 n_2 h_1^{-1} \in N.$$

So the multiplication of G is determined by N, H, and the action of H on N by conjugation. This leads to the following construction.

Definition 11.6.1 (Semidirect product)

Let N and H be groups and $\varphi : H \to \mathrm{Aut}(N)$ a homomorphism, that is, an action of H on N by automorphisms. The ***semidirect product*** $N \rtimes_\varphi H$ is the set $N \times H$ with the multiplication

$$(n_1, h_1) \cdot (n_2, h_2) = \bigl(n_1\, \varphi(h_1)(n_2),\ h_1 h_2\bigr).$$

This is a group. The element (e, e) is an identity, and $(\varphi(h^{-1})(n^{-1}), h^{-1})$ is an inverse of (n, h) on both sides. For associativity, using that $\varphi(h_1)$ is a homomorphism and $\varphi(h_1)\varphi(h_2) = \varphi(h_1h_2)$,

$$\begin{aligned}\big((n_1, h_1)(n_2, h_2)\big)(n_3, h_3) &= \big(n_1\,\varphi(h_1)(n_2)\,\varphi(h_1h_2)(n_3),\ h_1h_2h_3\big)\\ &= (n_1, h_1)\big((n_2, h_2)(n_3, h_3)\big).\end{aligned}$$

The map $(n, h) \mapsto h$ is a homomorphism onto H with kernel $N \times \{e\} \cong N$, so $N \times \{e\}$ is normal; $\{e\} \times H$ is a subgroup isomorphic to H; and $(e, h)(n, e)(e, h)^{-1} = (\varphi(h)(n), e)$, so inside $N \rtimes_\varphi H$ the action φ becomes conjugation. When φ is trivial ($\varphi(h) = \mathrm{id}_N$ for all h), the semidirect product is the direct product $N \times H$.

Proposition 11.6.2 (Recognizing semidirect products)

Let $N, H \leqslant G$ with N normal, and suppose every element of G can be written uniquely as $g = nh$ with $n \in N$ and $h \in H$. Then $G \cong N \rtimes_\varphi H$, where $\varphi(h)(n) = hnh^{-1}$.

Proof. Conjugation by h maps the normal subgroup N to itself, so $\varphi : H \to \mathrm{Aut}(N)$ is a homomorphism. The map $N \rtimes_\varphi H \to G$, $(n, h) \mapsto nh$, is a bijection by hypothesis, and it is a homomorphism, since

$$(n_1h_1)(n_2h_2) = n_1(h_1n_2h_1^{-1})h_1h_2 = \big(n_1\,\varphi(h_1)(n_2)\big)(h_1h_2). \qquad \square$$

Lemma 11.6.3 (Unique factorization)

Let $N, H \leqslant G$ with $N \cap H = \{e\}$. Then the products nh with $n \in N$, $h \in H$ are pairwise distinct. So if G is finite and $|G| = |N| \cdot |H|$, every element of G can be written uniquely as nh. The condition $N \cap H = \{e\}$ holds whenever $|N|$ and $|H|$ are coprime, because the order of $N \cap H$ divides both.

Proof. If $n_1h_1 = n_2h_2$, then $n_2^{-1}n_1 = h_2h_1^{-1} \in N \cap H = \{e\}$, so $n_1 = n_2$ and $h_1 = h_2$. If $|G| = |N||H|$, the $|N||H|$ distinct products fill G. $\square$

Corollary 11.6.4 (Direct product criterion)

Let G be a finite group with normal subgroups N and H such that $N \cap H = \{e\}$ and $|G| = |N| \cdot |H|$. Then $G \cong N \times H$.

Proof. By Lemma 11.6.3 and Proposition 11.6.2, $G \cong N \rtimes_\varphi H$ with $\varphi(h)(n) = hnh^{-1}$. For $n \in N$ and $h \in H$, the element $nhn^{-1}h^{-1}$ lies in N (as $hn^{-1}h^{-1} \in N$) and in H (as $nhn^{-1} \in H$), so it is e, and $hnh^{-1} = n$. So φ is trivial and $G \cong N \times H$. $\square$

Example 11.6.5 (Dihedral groups as semidirect products)

In S_3, take $N = A_3 = \langle \sigma \rangle$ with $\sigma = (1\,2\,3)$, and $H = \{e, \tau\}$ with $\tau = (1\,2)$. Then N is normal, $N \cap H = \{e\}$ and $|S_3| = 3 \cdot 2$, so $S_3 \cong \mathbb{Z}/3 \rtimes_\varphi \mathbb{Z}/2$, where the generator of $\mathbb{Z}/2$ acts by $\tau\sigma\tau^{-1} = \sigma^{-1}$, that is, by inversion. In the same way $D_n \cong \mathbb{Z}/n \rtimes_\varphi \mathbb{Z}/2$, with N the subgroup of rotations and $H = \{e, s\}$ for a reflection s, since $srs^{-1} = r^{-1}$. For $n \geqslant 3$ these groups are not abelian, so they are not direct products of the abelian groups $\mathbb{Z}/n$ and $\mathbb{Z}/2$.

11.6.2 Groups of Order 15 and 21

Example 11.6.6 (Groups of order 15)

Let $|G| = 15 = 3 \cdot 5$. By Sylow's third theorem, $s_3 \mid 5$ and $s_3 \equiv 1 \pmod 3$, so $s_3 = 1$; and $s_5 \mid 3$ and $s_5 \equiv 1 \pmod 5$, so $s_5 = 1$. So the Sylow 3-subgroup H and the Sylow 5-subgroup K are normal. Their orders are coprime and $|H||K| = 15$, so $G \cong H \times K \cong \mathbb{Z}/3 \times \mathbb{Z}/5$ by Corollary 11.6.4, and $\mathbb{Z}/3 \times \mathbb{Z}/5 \cong \mathbb{Z}/15$ because $(1,1)$ has order 15. Every group of order 15 is cyclic.

Example 11.6.7 (Groups of order 21)

Let $|G| = 21 = 3 \cdot 7$. Sylow gives $s_7 \mid 3$ and $s_7 \equiv 1 \pmod 7$, so $s_7 = 1$, and the Sylow 7-subgroup $K = \langle y \rangle \cong \mathbb{Z}/7$ is normal. Also $s_3 \mid 7$ and $s_3 \equiv 1 \pmod 3$, so $s_3 \in \{1, 7\}$. Let $H = \langle x \rangle$ be a Sylow 3-subgroup.

If $s_3 = 1$, then H is normal too, and $G \cong \mathbb{Z}/3 \times \mathbb{Z}/7 \cong \mathbb{Z}/21$ as for order 15.

If $s_3 = 7$, then $G \cong K \rtimes_\varphi H$ by Lemma 11.6.3 and Proposition 11.6.2, where $\varphi(x)$ is conjugation by x on K. An automorphism of K sends y to another generator, so $xyx^{-1} = y^a$ for some $a \in \{1, \dots, 6\}$. Since $x^3 = e$, $y = x^3 y x^{-3} = y^{a^3}$, so $a^3 \equiv 1 \pmod 7$, and $a \in \{1, 2, 4\}$. If $a = 1$, then x and y commute, G is abelian, and every subgroup is normal, contradicting $s_3 = 7$. If $a = 4$, replace the generator x of H by x^2: then $x^2 y x^{-2} = y^{16} = y^2$. So we may take $a = 2$, and

$$G = \langle x, y \rangle, \qquad x^3 = y^7 = e, \qquad xyx^{-1} = y^2,$$

where every element is uniquely $y^i x^j$ $(0 \leqslant i < 7,\ 0 \leqslant j < 3)$ and these relations determine all products. Conversely, the semidirect product $\mathbb{Z}/7 \rtimes_\varphi \mathbb{Z}/3$, in which $1 \in \mathbb{Z}/3$ acts on $\mathbb{Z}/7$ by multiplication by 2 (an automorphism of order 3, since $2^3 = 8 \equiv 1 \pmod 7$), is a non-abelian group of order 21. So there are exactly two groups of order 21 up to isomorphism: $\mathbb{Z}/21$ and $\mathbb{Z}/7 \rtimes \mathbb{Z}/3$.

11.6.3 Groups of Order 12

Proposition 11.6.8 (Groups of order 12)
Up to isomorphism there are exactly five groups of order 12:

$$\mathbb{Z}/12, \qquad \mathbb{Z}/2 \times \mathbb{Z}/6, \qquad A_4, \qquad D_6, \qquad \mathbb{Z}/3 \rtimes \mathbb{Z}/4,$$

where in the last group the generator of $\mathbb{Z}/4$ acts on $\mathbb{Z}/3$ by inversion.

Proof. Let $|G| = 12 = 2^2 \cdot 3$. By Sylow's theorems, the number s_2 of Sylow 2-subgroups (of order 4) divides 3 and is odd, so $s_2 \in \{1, 3\}$; and the number s_3 of Sylow 3-subgroups (of order 3) divides 4 and is $\equiv 1 \pmod 3$, so $s_3 \in \{1, 4\}$. Let H be a Sylow 2-subgroup and K a Sylow 3-subgroup. By Corollary 11.2.5, $H \cong \mathbb{Z}/4$ or $H \cong \mathbb{Z}/2 \times \mathbb{Z}/2$. Since $|H|$ and $|K|$ are coprime and $|H||K| = 12$, every element of G is uniquely a product kh with $k \in K$, $h \in H$ (Lemma 11.6.3).

At least one of H, K is normal. Suppose $s_3 = 4$, with Sylow 3-subgroups $K_1, \dots, K_4$. For $i \neq j$, $K_i \cap K_j$ is a proper subgroup of a group of prime order, so $K_i \cap K_j = \{e\}$, and the K_i contain $4 \cdot 2 = 8$ elements of order 3. A Sylow 2-subgroup contains no element of order 3, so it lies in the set of the remaining $12 - 8 = 4$ elements, and therefore equals it. Hence $s_2 = 1$.

Both normal. Then $G \cong K \times H$ by Corollary 11.6.4. With $H \cong \mathbb{Z}/4$ this is $\mathbb{Z}/3 \times \mathbb{Z}/4 \cong \mathbb{Z}/12$, and with $H \cong \mathbb{Z}/2 \times \mathbb{Z}/2$ it is $\mathbb{Z}/2 \times \mathbb{Z}/2 \times \mathbb{Z}/3 \cong \mathbb{Z}/2 \times \mathbb{Z}/6$.

K not normal ($s_3 = 4$). Let G act by conjugation on the set $\{K_1, K_2, K_3, K_4\}$ of Sylow 3-subgroups, giving a homomorphism $\psi : G \to S_4$. By Proposition 11.5.7, $|N(K_i)| = 12/4 = 3$, so $N(K_i) = K_i$. Let y_i generate K_i. Then y_i fixes K_i, and fixes no K_j with $j \neq i$, since $y_i \in N(K_j) = K_j$ would put y_i in $K_i \cap K_j = \{e\}$. As $y_i^3 = e$, $\psi(y_i)$ is a 3-cycle fixing K_i, and $\psi(y_i^2)$ is the other 3-cycle fixing K_i. So the image of ψ contains all eight 3-cycles, hence contains A_4 by Lemma 11.4.8. Since $|\psi(G)| \leqslant |G| = 12 = |A_4|$, ψ is injective with image A_4, and $G \cong A_4$.

K normal, H not normal ($s_3 = 1$, $s_2 = 3$). Write $K = \{e, y, y^2\}$. By Proposition 11.6.2, $G \cong K \rtimes_\varphi H$, where $\varphi : H \to \mathrm{Aut}(K)$ is conjugation. The group $\mathrm{Aut}(K) = \{\mathrm{id}, \iota\}$, where $\iota(y) = y^2 = y^{-1}$, has order 2, and φ is nontrivial: otherwise H would commute with K, and $G \cong K \times H$ would be abelian, so H would be normal.

- If $H = \langle x \rangle \cong \mathbb{Z}/4$, then $\varphi(x) = \iota$, so $xyx^{-1} = y^{-1}$, and $G \cong \mathbb{Z}/3 \rtimes \mathbb{Z}/4$ with the generator of $\mathbb{Z}/4$ acting by inversion.
- If $H \cong \mathbb{Z}/2 \times \mathbb{Z}/2$, then $\ker \varphi$ has order 2; let z generate it and let $x \in H \setminus \ker \varphi$. Then $H = \{e, x, z, xz\}$ and

$$x^2 = z^2 = y^3 = e, \qquad xz = zx, \qquad zy = yz, \qquad xyx^{-1} = y^{-1}.$$

Every element of G is uniquely $y^i z^j x^k$ ($0 \leqslant i < 3$; $j, k \in \{0, 1\}$), and these relations determine all products. In D_6, the symmetry group of the regular hexagon with rotation r by $\pi/3$ and a reflection s, the elements $y = r^2$, $z = r^3$ and $x = s$ satisfy the same relations (z commutes with y as both are rotations, and $sr^3 = r^{-3}s = r^3 s$), and they generate D_6, since $r = r^3 r^{-2}$. So $G \cong D_6$.

These five groups are pairwise non-isomorphic. The first two are abelian, and only $\mathbb{Z}/12$ has an element of order 12. Among the non-abelian ones, $\mathbb{Z}/3 \rtimes \mathbb{Z}/4$ has an element of order 4 while A_4 and D_6 do not, and D_6 has an element of order 6 while A_4 does not. Finally all five exist: the semidirect product $\mathbb{Z}/3 \rtimes \mathbb{Z}/4$ is defined for any homomorphism $\mathbb{Z}/4 \to \mathrm{Aut}(\mathbb{Z}/3)$, and it is non-abelian when the generator acts by inversion. □

NOTE The group D_6 is also isomorphic to $S_3 \times \mathbb{Z}/2$. In D_6, the subgroup $\langle r^2, s\rangle$ of order 6 is isomorphic to $D_3 \cong S_3$ and normal (it has index 2), the subgroup $\langle r^3\rangle \cong \mathbb{Z}/2$ is normal because r^3 commutes with every element, the two meet only in e, and $6 \cdot 2 = 12$; so Corollary 11.6.4 applies. Likewise $A_4 \cong V_4 \rtimes \mathbb{Z}/3$, with $\mathbb{Z}/3 = \langle (1\,2\,3)\rangle$ permuting the three non-identity elements of V_4 cyclically by conjugation.

Exercise 11.6.9 (Group actions, conjugacy, and the Sylow theorems)

(1) Let G be a group of order pq, where $p < q$ are primes and $p \nmid q - 1$. Prove that G is cyclic.

(2) Compute the class equation of S_5. Use it to determine all normal subgroups of S_5.

(3) Use Burnside's lemma to count the necklaces made from 6 beads in 2 colors, where two necklaces are the same if one can be rotated to obtain the other.

(4) Let G be a finite group acting on a finite set S, and let N be the number of orbits. Prove that $\sum_{s \in S} |\mathrm{Stab}(s)| = N \cdot |G|$.

(5) Compute the class equation of A_6, and use it to prove that A_6 is simple.

Solution. **(1)** The number s_q of Sylow q-subgroups divides p and is $\equiv 1 \pmod q$; since $p < q$, $s_q = 1$. The number s_p divides q and is $\equiv 1 \pmod p$, so $s_p \in \{1, q\}$; and $s_p = q$ is impossible, since $q \not\equiv 1 \pmod p$. So $s_p = s_q = 1$, and the Sylow subgroups $H \cong \mathbb{Z}/p$ and $K \cong \mathbb{Z}/q$ are both normal. Their orders are coprime with product pq, so $G \cong \mathbb{Z}/p \times \mathbb{Z}/q$ by

Corollary 11.6.4, and this group is cyclic, generated by $(1,1)$, which has order pq.

(2) The conjugacy classes of S_5 correspond to the partitions of 5. Counting each cycle type: the identity (1); transpositions, $\binom{5}{2} = 10$; products of two disjoint transpositions, $5 \cdot 3 = 15$ (choose the fixed point, then pair up the other four); 3-cycles, $\binom{5}{3} \cdot 2 = 20$; products of a 3-cycle and a disjoint transposition, $\binom{5}{3} \cdot 2 = 20$; 4-cycles, $5 \cdot 3! = 30$; 5-cycles, $4! = 24$. The class equation is

$$120 = 1 + 10 + 15 + 20 + 20 + 30 + 24.$$

A normal subgroup is a union of classes containing $\{e\}$ whose size divides 120. The proper divisors of 120 are at most 60, and the only sums $1 + s$ with s a sum of distinct terms of $10, 15, 20, 20, 30, 24$ that are divisors of 120 other than 1 and 120 are

$$1 + 15 + 24 = 40 \qquad \text{and} \qquad 1 + 15 + 20 + 24 = 60,$$

where in the second the 20 can be either the 3-cycles or the products of a 3-cycle and a transposition. Each of these three unions contains the 5-cycle $(1\,2\,3\,4\,5)$ and the element $(1\,2)(3\,4)$, and composing from right to left,

$$(1\,2\,3\,4\,5)(1\,2)(3\,4) = (1\,3\,5),$$

a 3-cycle. So the union of size 40 and the union of size 60 that uses the products of a 3-cycle and a transposition are not closed under multiplication. The remaining union, of the identity, the products of two transpositions, the 3-cycles and the 5-cycles, is the set of even permutations, A_5. So the normal subgroups of S_5 are $\{e\}$, A_5 and S_5, as Corollary 11.4.10 also shows.

(3) The rotation group is $G = \mathbb{Z}/6$, acting on the set S of 2-colorings of the 6 positions, $|S| = 2^6 = 64$. The rotation by k positions permutes the positions in $\gcd(k,6)$ cycles, and a coloring is fixed if and only if it is constant on each cycle, so it has $2^{\gcd(k,6)}$ fixed points. For $k = 0, 1, 2, 3, 4, 5$ this gives $64, 2, 4, 8, 4, 2$. By Burnside's lemma the number of necklaces is $\frac{1}{6}(64 + 2 + 4 + 8 + 4 + 2) = \frac{84}{6} = 14$.

(4) This is the double count in the proof of Burnside's lemma: $\sum_{s \in S} |\mathrm{Stab}(s)|$ and $\sum_{g \in G} |S^g|$ both count the pairs (g, s) with $g \cdot s = s$, and the second sum is $N \cdot |G|$ by Burnside's lemma.

(5) The even cycle types in S_6, counting fixed points as 1-cycles, and the sizes of their classes in S_6 are: the identity (1); two disjoint transpositions, $\binom{6}{4} \cdot 3 = 45$; 3-cycles, $\binom{6}{3} \cdot 2 = 40$; two disjoint 3-cycles, $\frac{1}{2}\binom{6}{3} \cdot 2 \cdot 2 = 40$; a 4-cycle and a disjoint transposition, $\binom{6}{4} \cdot 3! = 90$; 5-cycles, $6 \cdot 4! = 144$. These

add up to $360 = |A_6|$. By Proposition 11.4.4, only the class of 5-cycles (cycle lengths $5, 1$) splits in A_6, into two classes of 72. The class equation of A_6 is

$$360 = 1 + 45 + 40 + 40 + 90 + 72 + 72.$$

A normal subgroup $H \neq \{e\}$ of A_6 has $|H| = 1 + s$, with s a nonzero sum of distinct terms of $45, 40, 40, 90, 72, 72$, and $|H|$ divides 360. Since $s \geqslant 40$, $|H| \geqslant 41$, and the divisors of 360 between 41 and 359 are $45, 60, 72, 90, 120, 180$; so if $|H| \neq 360$, then $s \in \{44, 59, 71, 89, 119, 179\}$. The number 44 is not such a sum, since no term equals 44 and any two terms add up to at least 80. The others are odd, so they would have to use the only odd term 45, leaving $14, 26, 44, 74, 134$ as sums of distinct terms of $40, 40, 90, 72, 72$; but the sums of one or two of these terms are $40, 72, 90, 80, 112, 130, 144, 162$, and sums of three or more are at least 152. So no such s exists, $|H| = 360$, and A_6 is simple.

CHAPTER 12

Generators, Presentations, and Cayley Graphs

Section 3.6 introduced free groups (Definition 3.6.6), presentations $\langle S \mid R \rangle = F(S)/\langle\langle R \rangle\rangle$ (Definition 3.6.13), and the standard method for proving that a group has a given presentation: check that the relations hold in the group, which gives a surjection from the presented group (Proposition 3.6.14), and then show that this surjection is injective. It also posed the word problem: given two words in the generators of $\langle S \mid R \rangle$, decide whether they represent the same element. For groups whose elements we can compute with directly, such as groups of permutations or of matrices, this is easy: evaluate both words and compare. For a group known only through a presentation it can be hard, and by the Novikov–Boone theorem there are finitely presented groups for which no algorithm solves it.

This chapter develops two tools for understanding a group given by generators and relations. The *Cayley graph* turns the group into a geometric object, and *normal forms* single out a preferred word for each element. We apply them to the symmetric groups, the braid groups, and the groups $\mathrm{SL}_2(\mathbb{Z})$ and $\mathrm{PSL}_2(\mathbb{Z})$.

12.1 Cayley Graphs

12.1.1 Definition and First Examples

Definition 12.1.1 (Cayley graph)

Let G be a group generated by $g_1, \ldots, g_n$. The ***Cayley graph*** of G with respect to these generators has:

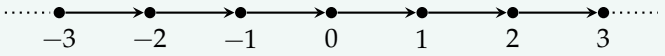
- one vertex for each element of G;
- for each $s \in G$ and each generator g_i, a directed edge from s to sg_i, labeled g_i.

Every vertex has one outgoing and one incoming edge with each label. When a generator satisfies $g_i = g_i^{-1}$, the edges $s \to sg_i$ and $sg_i \to s$ come in pairs, and we draw each pair as a single undirected edge. A word $x_1x_2\cdots x_k$ in the letters $g_i^{\pm 1}$ describes a path from e to $x_1\cdots x_k$, following the edge labeled g_i forward for each letter g_i and backward for each letter g_i^{-1}. Since the g_i generate G, the Cayley graph is connected.

NOTE G acts on its Cayley graph by left multiplication: $g \in G$ sends the vertex s to gs and the edge $s \to sg_i$ to the edge $gs \to gsg_i$, which has the same label. This action is transitive on vertices, and transitive on the edges with a given label, so the Cayley graph looks the same from every vertex.

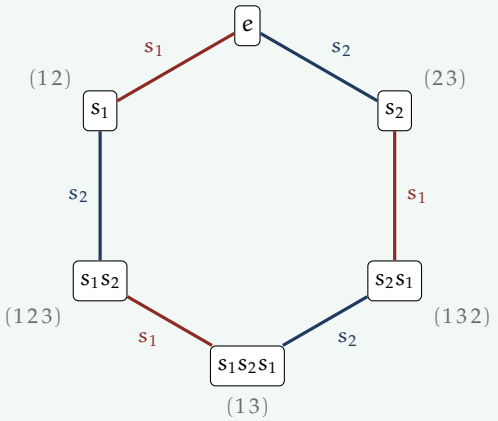
Example 12.1.2 (Cayley graph of $\mathbb{Z}$)

With the generator 1, the Cayley graph of $\mathbb{Z}$ is the integer number line:



Example 12.1.3 (Cayley graph of S_3)

Take the generators $s_1 = (1\,2)$ and $s_2 = (2\,3)$. Since $s_i^{-1} = s_i$, we draw undirected edges. The six vertices form a hexagon whose edges are alternately labeled s_1 (red) and s_2 (blue); next to each vertex we give the permutation in cycle notation.

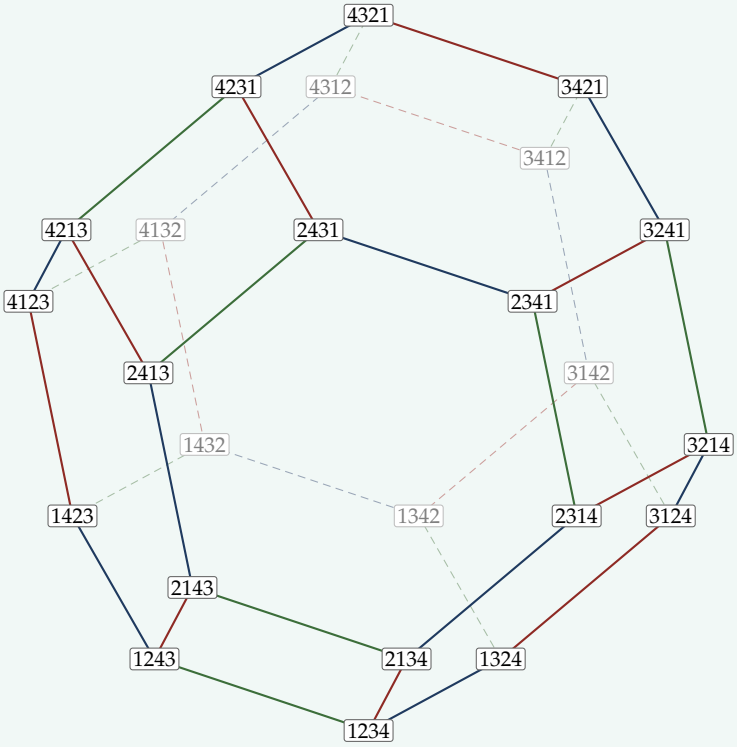


Every vertex has exactly one s_1-edge and one s_2-edge. The two paths from e to the bottom vertex show that $s_1s_2s_1 = s_2s_1s_2$; given $s_1^2 = s_2^2 = e$, this relation is equivalent to $(s_1s_2)^3 = e$, which says that the path $s_1s_2s_1s_2s_1s_2$

around the hexagon closes up. In fact $S_3 \cong \langle s_1, s_2 \mid s_1^2, s_2^2, (s_1 s_2)^3 \rangle$, the case $n = 3$ of Theorem 3.6.18.

Example 12.1.4 (Cayley graph of S_4)

Take the generators $s_1 = (1\,2)$, $s_2 = (2\,3)$, and $s_3 = (3\,4)$. In the figure, each vertex σ is labeled by the string $\sigma(1)\sigma(2)\sigma(3)\sigma(4)$ (the ***one-line notation*** of σ). The vertex σs_i is obtained from σ by swapping the entries in positions i and $i+1$, since $(\sigma s_i)(i) = \sigma(i+1)$ and $(\sigma s_i)(i+1) = \sigma(i)$. Edges labeled s_1, s_2, s_3 are drawn in red, blue, and green, and edges at the back are dashed.



The 24 vertices and 36 edges form the edges of a truncated octahedron, also called the ***permutohedron***. Its faces correspond to relations. The six squares have edges alternately labeled s_1 and s_3, and express $s_1 s_3 = s_3 s_1$. The eight hexagons have edges alternately labeled s_1, s_2 or s_2, s_3, and express $s_1 s_2 s_1 = s_2 s_1 s_2$ and $s_2 s_3 s_2 = s_3 s_2 s_3$. The identity 1234 is at the bottom and 4321 at the top; Section 12.2 will show that every edge going up increases the distance from e by one.

12.1.2 Word Length and Growth

Definition 12.1.5 (Word length)

Let G be generated by $g_1, \dots, g_n$. The ***word length*** $\ell(g)$ of $g \in G$ is the distance from e to g in the Cayley graph, where edges may be followed in either direction. Equivalently, $\ell(g)$ is the smallest k such that $g = x_1 x_2 \cdots x_k$ with each $x_j \in \{g_1^{\pm 1}, \dots, g_n^{\pm 1}\}$.

For finite groups the Cayley graph is finite, but word lengths remain interesting; we study them for S_n in the next section. For infinite groups, the Cayley graph also records how fast the group grows.

Definition 12.1.6 (Growth)

Let G be generated by $g_1, \dots, g_n$, and let $B(N)$ be the number of elements $g \in G$ with $\ell(g) \leqslant N$. The group has ***polynomial growth*** if there are constants C and d with $B(N) \leqslant C N^d$ for all $N \geqslant 1$, and ***exponential growth*** if there are constants $C > 0$ and $c > 1$ with $B(N) \geqslant C c^N$ for all N.

These notions do not depend on the choice of generators. If $g_1', \dots, g_m'$ is another generating set, with word length ℓ' and counting function B', choose L such that each $g_j'^{\pm 1}$ is a product of at most L letters $g_i^{\pm 1}$. Then $\ell(g) \leqslant L\ell'(g)$, so $B'(N) \leqslant B(LN)$, and a bound $B(N) \leqslant C N^d$ gives $B'(N) \leqslant C L^d N^d$. Exchanging the roles of the two generating sets, with a constant L', gives $B(M) \leqslant B'(L'M)$ for all M; so if $B(N) \geqslant C c^N$, then taking $M = \lfloor N/L' \rfloor$,

$$B'(N) \geqslant B(\lfloor N/L' \rfloor) \geqslant C c^{N/L' - 1} = \tfrac{C}{c}\left(c^{1/L'}\right)^N.$$

Example 12.1.7 (Growth of abelian and free groups)

If G is abelian and generated by $g_1, \dots, g_n$, every element of word length at most N can be written $m_1 g_1 + \cdots + m_n g_n$ with $|m_1| + \cdots + |m_n| \leqslant N$, so $B(N) \leqslant (2N+1)^n \leqslant 3^n N^n$: finitely generated abelian groups have polynomial growth. In the free group F_k with $k \geqslant 2$, distinct reduced words are distinct elements, and the word length of an element is the length of its reduced word. There are $2k(2k-1)^{N-1} \geqslant 3^N$ reduced words of length exactly $N \geqslant 1$, so F_k has exponential growth. (The free group $F_1 \cong \mathbb{Z}$ has polynomial growth.)

NOTE A deep theorem of Gromov (1981) characterizes polynomial growth: a finitely generated group has polynomial growth if and only if it has a nilpotent subgroup of finite index. Here a group G is *nilpotent* if the sequence $G_1 = G$, $G_{j+1} =$ the subgroup generated by all $ghg^{-1}h^{-1}$ with $g \in G$ and $h \in G_j$, reaches $\{e\}$; abelian groups are the case $G_2 = \{e\}$.

12.2 Permutation Diagrams and Normal Forms in S_n

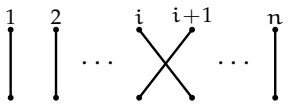
Throughout this section, $s_i = (i\ \ i{+}1)$ for $1 \leqslant i \leqslant n-1$ denotes an adjacent transposition in S_n, and ℓ is the word length with respect to $s_1, \dots, s_{n-1}$. Since $s_i^{-1} = s_i$, words in the $s_i^{\pm 1}$ are just words in the s_i.

12.2.1 Permutation Diagrams

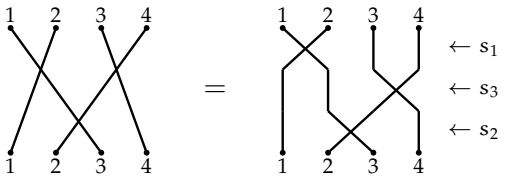
Definition 12.2.1 (Permutation diagram)

The ***diagram*** of $\sigma \in S_n$ consists of points $1, \dots, n$ on a top line, points $1, \dots, n$ on a bottom line, and for each i a strand joining the top point i to the bottom point $\sigma(i)$. The strands may be drawn as straight segments, or as any curves that move steadily downward; only which point is joined to which matters.

The diagram of a product $\sigma\tau$ is obtained by placing the diagram of τ above the diagram of σ: the strand starting at i reaches $\tau(i)$ in the middle and $\sigma(\tau(i))$ at the bottom. This matches our convention that $\sigma\tau$ applies τ first. The diagram of s_i is a single crossing:



So writing $\sigma = s_{i_1} s_{i_2} \cdots s_{i_k}$ amounts to drawing a diagram of σ in k layers, each containing a single crossing, with s_{i_k} at the top and s_{i_1} at the bottom. For example, here is the permutation σ with one-line notation 3142 (so $\sigma(1) = 3$, $\sigma(2) = 1$, $\sigma(3) = 4$, $\sigma(4) = 2$; in cycle notation $\sigma = (1\,3\,4\,2)$), drawn with straight strands and in layers:



The layered diagram shows that $\sigma = s_2 s_3 s_1$. The straight diagram has three crossings, one for each pair of strands whose order is reversed.

12.2.2 Word Length Equals the Number of Inversions

Recall from Definition 3.4.10 that an *inversion* of $\sigma \in S_n$ is a pair (i, j) with $1 \leqslant i < j \leqslant n$ and $\sigma(i) > \sigma(j)$, that $\mathrm{inv}(\sigma)$ is the number of inversions, and that $\mathrm{sgn}(\sigma) = (-1)^{\mathrm{inv}(\sigma)}$. In the straight diagram of σ, the strands starting at $i < j$ cross exactly when $\sigma(i) > \sigma(j)$, so $\mathrm{inv}(\sigma)$ is the number of crossings.

Lemma 12.2.2 (Multiplying by s_i)

Let $\sigma \in S_n$ and $1 \leqslant i \leqslant n-1$. Then

$$\mathrm{inv}(\sigma s_i) = \begin{cases} \mathrm{inv}(\sigma) + 1 & \text{if } \sigma(i) < \sigma(i+1), \\ \mathrm{inv}(\sigma) - 1 & \text{if } \sigma(i) > \sigma(i+1). \end{cases}$$

Proof. The permutation σs_i takes the same values as σ, except that the values in positions i and $i+1$ are swapped: $(\sigma s_i)(p) = \sigma(s_i(p))$. Let $p < q$ be positions with $(p, q) \neq (i, i+1)$. Since s_i only exchanges the adjacent positions i and $i+1$, we still have $s_i(p) < s_i(q)$, and (p, q) is an inversion of σs_i if and only if $(s_i(p), s_i(q))$ is an inversion of σ. The map $(p, q) \mapsto (s_i(p), s_i(q))$ is a bijection of the set of pairs other than $(i, i+1)$, so σ and σs_i have the same number of inversions among these pairs. Finally, $(i, i+1)$ is an inversion of σs_i if and only if $\sigma(i+1) > \sigma(i)$, that is, if and only if it is not an inversion of σ. □

Theorem 12.2.3 (Word length in S_n)

For every $\sigma \in S_n$, the word length of σ with respect to $s_1, \ldots, s_{n-1}$ is $\ell(\sigma) = \mathrm{inv}(\sigma)$.

Proof. First, $\mathrm{inv}(\sigma) \leqslant k$ whenever $\sigma = s_{i_1} \cdots s_{i_k}$: by Lemma 12.2.2, each additional factor changes the number of inversions by ± 1, and $\mathrm{inv}(e) = 0$.

Next we show by induction on $\mathrm{inv}(\sigma)$ that σ is a product of $\mathrm{inv}(\sigma)$ generators. If $\mathrm{inv}(\sigma) = 0$, then $\sigma(1) < \sigma(2) < \cdots < \sigma(n)$, so $\sigma = e$, the empty product. If $\mathrm{inv}(\sigma) > 0$, then $\sigma(i) > \sigma(i+1)$ for some i, since otherwise σ would be increasing. By Lemma 12.2.2, $\mathrm{inv}(\sigma s_i) = \mathrm{inv}(\sigma) - 1$, so by induction σs_i is a product of $\mathrm{inv}(\sigma) - 1$ generators, and $\sigma = (\sigma s_i) s_i$ is a product of $\mathrm{inv}(\sigma)$ of them. □

In the language of diagrams: every layered diagram of σ has a crossing for each pair of strands whose order is reversed, and if the strands starting at adjacent points i and $i+1$ cross, we can draw that crossing first, as the top layer s_i, and continue with a diagram of σs_i.

Corollary 12.2.4 (The longest element)
The permutation

$$w_0 = \begin{pmatrix} 1 & 2 & \cdots & n \\ n & n-1 & \cdots & 1 \end{pmatrix},$$

with $w_0(i) = n+1-i$, has word length $\binom{n}{2} = \frac{n(n-1)}{2}$, and every other element of S_n has smaller word length.

Proof. Every pair $i < j$ is an inversion of w_0, so $\mathrm{inv}(w_0) = \binom{n}{2}$. Any $\sigma \neq w_0$ is not decreasing, so it has a pair $i < j$ with $\sigma(i) < \sigma(j)$, and $\mathrm{inv}(\sigma) < \binom{n}{2}$. □

Proposition 12.2.5 (Shortest words ending in s_i)
Let $\sigma \in S_n$ and $1 \leqslant i \leqslant n-1$. The following are equivalent:

(a) some shortest word for σ ends with s_i;
(b) $\ell(\sigma s_i) < \ell(\sigma)$;
(c) $\sigma(i) > \sigma(i+1)$.

The set of such i is the ***ending set*** of σ.

Proof. (a)⇒(b): if $\sigma = us_i$ with u a word of length $\ell(\sigma)-1$, then $\sigma s_i = u$ has length at most $\ell(\sigma)-1$. (b)⇒(a): by Lemma 12.2.2 and Theorem 12.2.3, $\ell(\sigma s_i) = \ell(\sigma)-1$; a shortest word u for σs_i followed by s_i is a word for σ of length $\ell(\sigma)$ ending with s_i. (b)⇔(c) is Lemma 12.2.2, since $\ell = \mathrm{inv}$. □

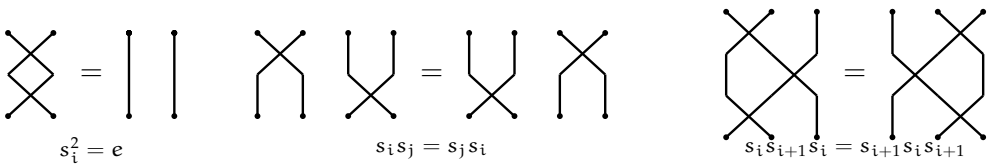
Similarly, some shortest word for σ begins with s_i if and only if i is in the ending set of σ^{-1} (reverse the word); these i form the ***starting set*** of σ.

12.2.3 Diagrams and the Presentation of S_n

By Theorem 3.6.18, the map $s_i \mapsto (i\ i{+}1)$ induces an isomorphism

$$\left\langle s_1, \ldots, s_{n-1} \;\middle|\; s_i^2 = e,\ \ s_i s_j = s_j s_i\ (|i-j| \geqslant 2),\ \ s_i s_{i+1} s_i = s_{i+1} s_i s_{i+1} \right\rangle \cong S_n.$$

In terms of diagrams, the three kinds of relations are three local moves:



two consecutive crossings of the same two strands cancel; crossings of strands far apart slide past each other; and a strand can be moved across the crossing of two

others. The presentation says that every identity between words in the s_i that holds in S_n follows from these three moves.

12.2.4 A Normal Form for S_n

Definition 12.2.6 (Normal form)

A ***normal form*** for a group $G = \langle S \mid R \rangle$ is a set of words in the letters $s^{\pm 1}$, $s \in S$, such that every element of G is represented by exactly one of these words.

If we have a normal form and an algorithm that rewrites any word as the normal-form word representing the same element, the word problem is solved: two words represent the same element if and only if they have the same normal form.

Proposition 12.2.5 lets us list all the shortest words for a permutation σ: for each i in the ending set of σ, find the shortest words for σs_i and append s_i. Making the same choice every time singles out one of them.

Definition 12.2.7 (Normal form for S_n)

For $\sigma \in S_n$, define a word $N(\sigma)$ in $s_1, \dots, s_{n-1}$ recursively: $N(e)$ is the empty word, and for $\sigma \neq e$,

$$N(\sigma) = N(\sigma s_i)\, s_i, \qquad \text{where } i \text{ is the smallest index with } \sigma(i) > \sigma(i+1).$$

Since $\mathrm{inv}(\sigma s_i) = \mathrm{inv}(\sigma) - 1$ at each step, the recursion stops after $\mathrm{inv}(\sigma)$ steps, and $N(\sigma)$ is a word of length $\mathrm{inv}(\sigma) = \ell(\sigma)$ representing σ: a shortest word for σ. As σ runs over S_n, the words $N(\sigma)$ form a normal form for S_n, with respect to the presentation of Theorem 3.6.18.

Example 12.2.8 (A normal form)

Let $\sigma = 3142$ as above. Its ending set is $\{1, 3\}$, since $3 > 1$ and $4 > 2$. Following the smallest index each time:

$$3142 \xrightarrow{s_1} 1342 \xrightarrow{s_3} 1324 \xrightarrow{s_2} 1234,$$

where each arrow multiplies on the right by s_i, that is, swaps the entries in positions i and $i+1$. So $N(\sigma) = s_2 s_3 s_1$, the word read off the layered diagram above. Following index 3 first instead, $3142 \xrightarrow{s_3} 3124 \xrightarrow{s_1} 1324 \xrightarrow{s_2} 1234$ gives

the other shortest word, $s_2s_1s_3$; since 1342, 3124 and 1324 each have a single index in their ending sets, these two are the only shortest words for σ.

For S_n, whose elements we can compute with directly, the word problem is easy anyway. The point is that the same method, describing an element by a preferred word built greedily from the end, also works in groups we cannot compute with so directly, such as the braid groups.

12.3 The Braid Group

12.3.1 Definition and Braid Diagrams

Definition 12.3.1 (Braid group)

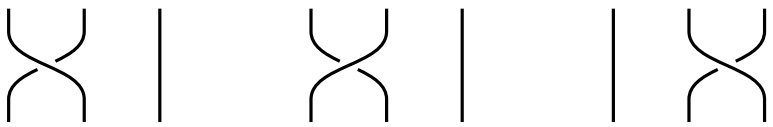
For $n \geqslant 1$, the ***braid group*** on n strands is

$$B_n = \left\langle s_1, \ldots, s_{n-1} \;\middle|\; s_i s_j = s_j s_i \ (|i-j| \geqslant 2),\ \ s_i s_{i+1} s_i = s_{i+1} s_i s_{i+1} \right\rangle.$$

This is the presentation of S_n in Theorem 3.6.18 without the relations $s_i^2 = e$. Two homomorphisms come straight from Proposition 3.6.14:

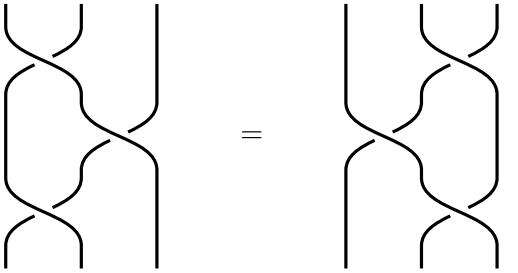
- $B_n \to S_n$, $s_i \mapsto (i\ \ i{+}1)$, which is surjective because the adjacent transpositions generate S_n. Its kernel is called the ***pure braid group*** PB_n.
- $B_n \to \mathbb{Z}$, $s_i \mapsto 1$: both sides of each relation have the same number of letters. It sends s_1^k to k, so for $n \geqslant 2$ the generator s_1 has infinite order and B_n is infinite. In particular $s_i^2 \neq e$ in B_n.

Braids are best understood as pictures. A braid on n strands consists of n strands in space joining n points on a top line to n points on a bottom line, each strand moving steadily downward; in a drawing, one strand passes over the other at each crossing. Two braids are the same if one can be deformed into the other without moving the endpoints and without strands passing through each other (an *isotopy*). The generator s_i is the braid in which the strand at position i crosses over the strand at position $i+1$, and s_i^{-1} is the one in which it crosses under:



s_1 $\qquad$ s_1^{-1} $\qquad$ s_2

As for permutations, the picture of a product $\beta\gamma$ is the picture of γ placed above that of β. The relations of B_n hold for pictures: $s_i s_i^{-1} = e$ because the two crossings can be pulled apart; $s_i s_j = s_j s_i$ for $|i - j| \geqslant 2$ because crossings of strands far apart slide past each other; and the braid relation $s_1 s_2 s_1 = s_2 s_1 s_2$ is the isotopy below, in which the middle strand slides across the crossing of the other two.



$s_1 s_2 s_1$ $\qquad\qquad$ $s_2 s_1 s_2$

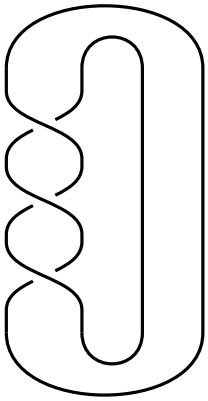
Conversely, one can show that every braid is isotopic to a product of the $s_i^{\pm 1}$, and that two products give isotopic braids only when they are equal in B_n; so B_n is exactly the group of braids on n strands. We will not prove this. The homomorphism $B_n \to S_n$ forgets which strand passes over which and records only where each strand ends, so PB_n consists of the braids in which every strand ends where it started.

12.3.2 Closures of Braids

Definition 12.3.2 (Closure of a braid)

The ***closure*** of a braid is obtained by joining each bottom endpoint to the top endpoint directly above it by an arc running around the right side of the picture, without creating new crossings.

The closure of a braid is a collection of closed curves in space: a *knot* if it is a single curve, and a *link* otherwise. For example, the closure of $s_1^3 \in B_2$ is the trefoil knot:



NOTE Markov's theorem connects braids to knot theory. Every knot or link in $\mathbb{R}^3$ is isotopic to the closure of a braid, and two braids have isotopic closures if and only if they are related by a sequence of moves of two kinds: conjugation, which replaces $\beta \in B_n$ by $g\beta g^{-1}$ with $g \in B_n$; and stabilization, which replaces $\beta \in B_n$ by $\beta s_n^{\pm 1} \in B_{n+1}$, where B_n sits inside B_{n+1} as the braids in which the last strand runs straight down.

12.3.3 Permutation Braids and the Garside Normal Form

B_n is much larger than S_n, but its algorithmic questions can be approached by the method of Section 12.2. We describe the result without proofs.

A ***permutation braid*** is a braid in which any two strands cross at most once and every crossing is positive, that is, of the kind in s_i rather than s_i^{-1}. For each $\sigma \in S_n$ there is exactly one permutation braid P_σ whose strands end at the positions given by σ: it is the product of the generators s_i along any shortest word for σ. So permutation braids are in bijection with S_n, and the starting and ending sets of P_σ are defined to be those of σ.

Let $\Delta = P_{w_0}$ be the longest permutation braid, where w_0 is the longest element of S_n (Corollary 12.2.4); in B_3, $\Delta = s_1 s_2 s_1$, pictured above. Every i lies in the starting set of w_0, so $\Delta = s_i P$ for some permutation braid P, and $s_i^{-1}\Delta = P$ is again a permutation braid. Hence each $s_i^{-1} = (s_i^{-1}\Delta)\Delta^{-1}$ is a permutation braid times Δ^{-1}. Moreover $s_i\Delta = \Delta s_{n-i}$ for every i, so conjugation by Δ permutes the generators, and powers of Δ can be moved to the left of any word. Consequently every $g \in B_n$ can be written as

$$g = \Delta^k P_1 P_2 \cdots P_r, \qquad k \in \mathbb{Z}, \quad P_1, \ldots, P_r \text{ permutation braids.}$$

Moving letters to the left whenever possible, we can arrange that for each j, the starting set of P_{j+1} is contained in the ending set of P_j. Otherwise the first letter

s_i of some shortest word for P_{j+1} could be moved to the end of P_j, and $P_j s_i$ would still be a permutation braid. With k as large as possible and no P_j trivial, this expression is unique. It is the ***Garside normal form*** (Garside, 1969; Thurston and Elrifai–Morton, early 1990s), and computing it solves the word problem in B_n.

12.4 $\mathrm{SL}_2(\mathbb{Z})$ and $\mathrm{PSL}_2(\mathbb{Z})$

12.4.1 Generators

Definition 12.4.1 ($\mathrm{SL}_2(\mathbb{Z})$ and $\mathrm{PSL}_2(\mathbb{Z})$)

The group

$$\mathrm{SL}_2(\mathbb{Z}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d \in \mathbb{Z},\ ad - bc = 1 \right\}$$

consists of the 2×2 integer matrices with determinant 1, under matrix multiplication. (It is closed under products since the determinant is multiplicative, and the inverse $\left(\begin{smallmatrix} d & -b \\ -c & a \end{smallmatrix}\right)$ has integer entries.) The subgroup $\{\pm I\}$ is normal, since $\pm I$ commutes with every matrix, and the quotient

$$\mathrm{PSL}_2(\mathbb{Z}) = \mathrm{SL}_2(\mathbb{Z}) / \{\pm I\}$$

is the ***modular group***. We denote the image of a matrix M in $\mathrm{PSL}_2(\mathbb{Z})$ again by M; thus $M = -M$ in $\mathrm{PSL}_2(\mathbb{Z})$.

Proposition 12.4.2 (Generators of $\mathrm{SL}_2(\mathbb{Z})$)

$\mathrm{SL}_2(\mathbb{Z})$ is generated by

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Consequently, $\mathrm{PSL}_2(\mathbb{Z})$ is generated by the images of S and T.

Proof. Let H be the subgroup generated by S and T. We show that every $M = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$ lies in H, by induction on $|c|$; the idea is to run the Euclidean algorithm on the first column.

If $c = 0$, then $ad = 1$, so $a = d = \pm 1$. Thus $M = \left(\begin{smallmatrix} 1 & b \\ 0 & 1 \end{smallmatrix}\right) = T^b$ or $M = \left(\begin{smallmatrix} -1 & b \\ 0 & -1 \end{smallmatrix}\right) = S^2 T^{-b}$, since $S^2 = -I$. Either way $M \in H$.

If $c \neq 0$, write $a = qc + r$ with $0 \leqslant r < |c|$. Then

$$T^{-q}M = \begin{pmatrix} 1 & -q \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} r & b - qd \\ c & d \end{pmatrix},$$

$$S^{-1}T^{-q}M = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} r & b - qd \\ c & d \end{pmatrix} = \begin{pmatrix} c & d \\ -r & qd - b \end{pmatrix}.$$

The lower left entry of $S^{-1}T^{-q}M$ is $-r$, and $|r| < |c|$, so by induction $S^{-1}T^{-q}M \in H$. Hence $M = T^q S(S^{-1}T^{-q}M) \in H$. Since $\mathrm{SL}_2(\mathbb{Z}) \to \mathrm{PSL}_2(\mathbb{Z})$ is surjective, the images of S and T generate $\mathrm{PSL}_2(\mathbb{Z})$. □

Two other pairs of generators are useful.

- S and $R = ST = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 1 \end{smallmatrix}\right)$, which have finite order: $S^2 = R^3 = -I$, so $S^4 = R^6 = I$ (Exercise 12.4.6(3)), and in $\mathrm{PSL}_2(\mathbb{Z})$ they have orders 2 and 3. They generate, since $T = S^{-1}R$.
- T and $T' = \left(\begin{smallmatrix} 1 & 0 \\ -1 & 1 \end{smallmatrix}\right) = (TST)^{-1} = STS^{-1}$, which are conjugate. A computation gives $TT' = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 1 \end{smallmatrix}\right)$, $(TT')^3 = -I$, and $TT'T = T'TT' = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right) = -S$. So $S = (TT')^3\, TT'T$, and T, T' generate $\mathrm{SL}_2(\mathbb{Z})$.

12.4.2 Presentations of $\mathrm{PSL}_2(\mathbb{Z})$ and $\mathrm{SL}_2(\mathbb{Z})$

Theorem 12.4.3 (Presentation of $\mathrm{PSL}_2(\mathbb{Z})$)
The map $s \mapsto S$, $r \mapsto R$ induces an isomorphism

$$\langle s, r \mid s^2, r^3 \rangle \cong \mathrm{PSL}_2(\mathbb{Z}).$$

Proof. Let $G = \langle s, r \mid s^2, r^3 \rangle$. Since $S^2 = R^3 = -I$, which is the identity of $\mathrm{PSL}_2(\mathbb{Z})$, Proposition 3.6.14 gives a homomorphism $\pi : G \to \mathrm{PSL}_2(\mathbb{Z})$ with $\pi(s) = S$ and $\pi(r) = R$, and it is surjective because S and R generate $\mathrm{PSL}_2(\mathbb{Z})$. We must show that π is injective.

Since $s^{-1} = s$ and $r^{-1} = r^2$, every element of G is a product of the letters s, r, r^{-1}. In such a product we may cancel ss and rr^{-1}, $r^{-1}r$, and replace rr by r^{-1} and $r^{-1}r^{-1}$ by r; each step shortens the word. So every element of G is represented by an ***alternating word***: a word in s, r, r^{-1} in which s alternates with $r^{\pm 1}$. It suffices to show that $\pi(w) \neq e$ for every nonempty alternating word w.

Suppose not, and let w be a nonempty alternating word with $\pi(w) = e$, of the smallest possible length. Note that $\pi(gwg^{-1}) = e$ whenever $\pi(w) = e$.

- w has length at least 2, because S, R, and R^{-1} are not $\pm I$.
- w does not begin and end with s: otherwise $w = sus$ with u nonempty, and $u = s^{-1}ws^{-1}$ would be a shorter alternating word with $\pi(u) = e$.

- w does not begin and end with letters r^ε and r^δ ($\varepsilon, \delta = \pm 1$): otherwise $w = r^\varepsilon u r^\delta$ with u beginning and ending with s, and $r^{-\varepsilon} w r^\varepsilon = u r^{\delta+\varepsilon}$ is either u (if $\delta = -\varepsilon$) or $ur^{\mp 1}$ (if $\delta = \varepsilon = \pm 1$, since $r^2 = r^{-1}$ and $r^{-2} = r$). Both are alternating words, shorter than w, on which π is trivial.
- If w begins with r^ε and ends with s, then $w = r^\varepsilon u$ and $r^{-\varepsilon} w r^\varepsilon = u r^\varepsilon$ is an alternating word of the same length that begins with s and ends with r^ε.

Replacing w by such a conjugate, we may therefore assume $w = sr^{\varepsilon_1} sr^{\varepsilon_2} \cdots sr^{\varepsilon_k}$ with $k \geqslant 1$ and each $\varepsilon_j = \pm 1$. Now compute in $SL_2(\mathbb{Z})$:

$$SR = \begin{pmatrix} -1 & -1 \\ 0 & -1 \end{pmatrix} = -\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad SR^{-1} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

So the matrix $SR^{\varepsilon_1} \cdots SR^{\varepsilon_k}$ is $\pm$ a product of k matrices, each equal to $A = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ or $B = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$. Let $M = \left(\begin{smallmatrix} p & q \\ x & y \end{smallmatrix}\right)$ have entries $\geqslant 0$ and determinant 1, so that neither column of M is zero. Then

$$MA = \begin{pmatrix} p & p+q \\ x & x+y \end{pmatrix}, \qquad MB = \begin{pmatrix} p+q & q \\ x+y & y \end{pmatrix}$$

again have entries $\geqslant 0$ and determinant 1, and their entry sums exceed that of M by $p + x \geqslant 1$ and $q + y \geqslant 1$ respectively. Starting from I, whose entry sum is 2, a product of $k \geqslant 1$ factors A or B therefore has nonnegative entries with sum at least $2 + k \geqslant 3$. It is not I (entry sum 2) and not $-I$ (negative entries). So $\pi(w) \neq e$, a contradiction. □

NOTE If $G_1 = \langle S_1 \mid R_1 \rangle$ and $G_2 = \langle S_2 \mid R_2 \rangle$ with S_1 and S_2 disjoint, the group $\langle S_1 \cup S_2 \mid R_1 \cup R_2 \rangle$ is called the ***free product*** $G_1 * G_2$. Theorem 12.4.3 says that $PSL_2(\mathbb{Z}) \cong \mathbb{Z}/2 * \mathbb{Z}/3$.

The presentation can be rewritten in terms of the other generators.

Corollary 12.4.4 (Other presentations)

With each generator sent to the matrix of the same name:

(a) $PSL_2(\mathbb{Z}) \cong \langle S, T \mid S^2, (ST)^3 \rangle$;

(b) $PSL_2(\mathbb{Z}) \cong \langle T, T' \mid (TT')^3,\ TT'T = T'TT' \rangle$;

(c) $SL_2(\mathbb{Z}) \cong \langle T, T' \mid (TT')^6,\ TT'T = T'TT' \rangle$.

Proof. We write s, r, t, u for the abstract generators and $G_1 = \langle s, r \mid s^2, r^3 \rangle$, $G_2 = \langle s, t \mid s^2, (st)^3 \rangle$, $G_3 = \langle t, u \mid (tu)^3, tut = utu \rangle$, $G_4 = \langle t, u \mid (tu)^6, tut = utu \rangle$. Each time, we define homomorphisms in both directions with Proposition 3.6.14

and check that both composites fix the generators, so that they are inverse isomorphisms.

(a) The assignments $s \mapsto s$, $r \mapsto st$ define $\alpha : G_1 \to G_2$, since $s^2 \mapsto s^2 = e$ and $r^3 \mapsto (st)^3 = e$. The assignments $s \mapsto s$, $t \mapsto s^{-1}r$ define $\beta : G_2 \to G_1$, since $(s \cdot s^{-1}r)^3 = r^3 = e$. Then $\beta\alpha(r) = ss^{-1}r = r$ and $\alpha\beta(t) = s^{-1}st = t$, so $G_1 \cong G_2$. Composing with Theorem 12.4.3, $t \mapsto S^{-1}R = T$.

(b) In G_3, let $\Delta = tut = utu$. Then $u\Delta = (utu)t = \Delta t$ and $t\Delta = (tut)u = \Delta u$, and $\Delta^2 = (tut)(utu) = (tu)^3 = e$. Define $\gamma : G_2 \to G_3$ by $s \mapsto \Delta$, $t \mapsto t$. The relations hold: $\Delta^2 = e$, and

$$(\Delta t)^3 = \Delta t \Delta\, t\Delta t = \Delta(\Delta u)(\Delta u)t = \Delta^2 u\Delta ut = \Delta^2 \Delta tut = \Delta^2\Delta\Delta = \Delta^4 = e,$$

using $t\Delta = \Delta u$ twice and $u\Delta = \Delta t$ once. Define $\delta : G_3 \to G_2$ by $t \mapsto t$, $u \mapsto sts^{-1} = sts$. In G_2, the relation $(st)^3 = e$ gives $tst = s^{-1}t^{-1}s^{-1} = st^{-1}s$, hence $tsts = st^{-1}$ and $tstst = s$. The relations of G_3 hold: $(t \cdot sts)^3 = (st^{-1})^3 = \big((ts)^3\big)^{-1} = \big(s^{-1}(st)^3 s\big)^{-1} = e$, and $t(sts)t = tstst = s$ while $(sts)t(sts) = (st)^3 s = s$. Finally $\delta\gamma(s) = \delta(tut) = tstst = s$, and $\gamma\delta(u) = \Delta t\Delta^{-1} = u\Delta\Delta^{-1} = u$. So $G_2 \cong G_3$, and under the isomorphism with $PSL_2(\mathbb{Z})$, $u = \delta(u) \mapsto STS^{-1} = T'$.

(c) The matrices satisfy the relations of G_4 (see the list of generators above), so Proposition 3.6.14 gives a homomorphism $\varphi : G_4 \to SL_2(\mathbb{Z})$, $t \mapsto T$, $u \mapsto T'$, which is surjective because T and T' generate. Let $z = (tu)^3 \in G_4$. As in (b), $\Delta = tut$ satisfies $\Delta t = u\Delta$, $\Delta u = t\Delta$, and $\Delta^2 = (tu)^3 = z$; so z commutes with t and u, hence is central, and $z^2 = (tu)^6 = e$. The subgroup $\langle z\rangle = \{e, z\}$ is therefore normal. The homomorphism $\psi : G_4 \to G_3$, $t \mapsto t$, $u \mapsto u$, kills z, so it induces $\bar{\psi} : G_4/\langle z\rangle \to G_3$; and $t \mapsto t\langle z\rangle$, $u \mapsto u\langle z\rangle$ defines an inverse $G_3 \to G_4/\langle z\rangle$, since $(tu)^3\langle z\rangle = z\langle z\rangle$ is trivial. So $\ker\psi = \langle z\rangle$. Now if $\varphi(g) = I$, then the image of g in $PSL_2(\mathbb{Z})$ is trivial, and by (b) this image is the image of $\psi(g)$ under the isomorphism $G_3 \cong PSL_2(\mathbb{Z})$; so $\psi(g) = e$ and $g \in \{e, z\}$. Since $\varphi(z) = (TT')^3 = -I \neq I$, we get $g = e$. So φ is injective. □

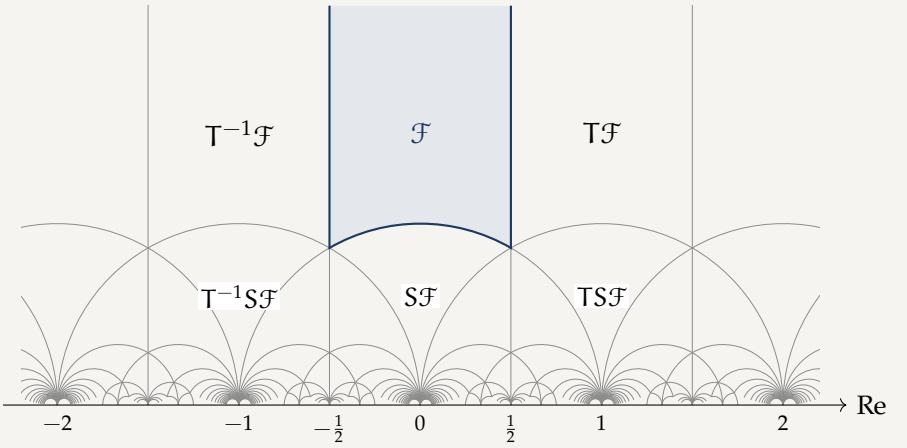
12.4.3 The Action on the Upper Half-Plane

NOTE Proposition 12.4.2 can also be seen geometrically. $SL_2(\mathbb{Z})$ acts on the upper half-plane $\mathbb{H} = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$ by Möbius transformations,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot z = \frac{az + b}{cz + d},$$

which preserve $\mathbb{H}$ because $\mathrm{Im}\,\frac{az+b}{cz+d} = \frac{\mathrm{Im}(z)}{|cz+d|^2}$. Since M and $-M$ act in the same way, this is an action of $PSL_2(\mathbb{Z})$. Here S acts by $z \mapsto -1/z$ and T by $z \mapsto z + 1$.

The region $\mathcal{F} = \{z \in \mathbb{H} : |z| \geqslant 1,\ |\operatorname{Re}(z)| \leqslant \tfrac{1}{2}\}$ is a *fundamental domain* for this action: its images under $\mathrm{PSL}_2(\mathbb{Z})$ tile $\mathbb{H}$, overlapping only along their edges. The figure shows $\mathcal{F}$ (shaded) and the edges of its images near it.



The tiles sharing an edge with $\mathcal{F}$ are $T\mathcal{F}$, $T^{-1}\mathcal{F}$, and $S\mathcal{F}$, so the tiles sharing an edge with $g\mathcal{F}$ are $gT\mathcal{F}$, $gT^{-1}\mathcal{F}$, and $gS\mathcal{F}$. Joining adjacent tiles therefore reproduces the Cayley graph of $\mathrm{PSL}_2(\mathbb{Z})$ with respect to S and T (with an undirected edge for $S = S^{-1}$), and the fact that every tile can be reached from $\mathcal{F}$ by crossing finitely many edges is equivalent to S and T generating $\mathrm{PSL}_2(\mathbb{Z})$.

12.4.4 Braids and the Modular Group

The braid relation $TT'T = T'TT'$ connects the two halves of this chapter.

Proposition 12.4.5 (B_3 and $\mathrm{PSL}_2(\mathbb{Z})$)

The assignment $s_1 \mapsto T$, $s_2 \mapsto T'$ defines a surjective homomorphism $B_3 \to \mathrm{PSL}_2(\mathbb{Z})$. Its kernel is generated by

$$\Delta^2 = (s_1 s_2 s_1)^2 = (s_1 s_2)^3,$$

which has infinite order, and this kernel is the center $Z(B_3)$. So there is an exact sequence

$$1 \to Z(B_3) \cong \mathbb{Z} \to B_3 \to \mathrm{PSL}_2(\mathbb{Z}) \to 1.$$

Proof. The relation $TT'T = T'TT'$ holds, so Proposition 3.6.14 gives the homomorphism, and it is surjective because T and T' generate. Since $B_3 = \langle s_1, s_2 \mid s_1 s_2 s_1 =$

$s_2s_1s_2\rangle$, the computations of part (b) of the proof of Corollary 12.4.4, which used only the braid relation, apply with $t = s_1$, $u = s_2$: $\Delta = s_1s_2s_1$ satisfies $\Delta s_1 = s_2\Delta$, $\Delta s_2 = s_1\Delta$, and $\Delta^2 = (s_1s_2)^3$. Hence Δ^2 commutes with s_1 and s_2, and $K = \langle\Delta^2\rangle$ is a central, hence normal, subgroup. The homomorphism $B_3 \to \mathbb{Z}$, $s_i \mapsto 1$, sends Δ^2 to 6, so Δ^2 has infinite order and $K \cong \mathbb{Z}$.

The kernel is K. As in part (c) of that proof, the homomorphism $\psi : B_3 \to G_3 = \langle t, u \mid (tu)^3, tut = utu\rangle$, $s_1 \mapsto t$, $s_2 \mapsto u$, kills Δ^2, and $t \mapsto s_1K$, $u \mapsto s_2K$ defines an inverse of the induced map $B_3/K \to G_3$; so $\ker\psi = K$. Our homomorphism is ψ followed by the isomorphism $G_3 \cong PSL_2(\mathbb{Z})$ of Corollary 12.4.4(b), so its kernel is K.

The kernel is the center. We have seen $K \subseteq Z(B_3)$. Conversely, a surjective homomorphism maps central elements to central elements, so it suffices to show that $PSL_2(\mathbb{Z})$ has trivial center. Suppose $M = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in SL_2(\mathbb{Z})$ satisfies $MT = \pm TM$ and $MS = \pm SM$. Comparing

$$MT = \begin{pmatrix} a & a+b \\ c & c+d \end{pmatrix} \quad \text{and} \quad TM = \begin{pmatrix} a+c & b+d \\ c & d \end{pmatrix}:$$

the sign $-$ would force $c = -c$ and $a = -a - c$, so $a = c = 0$, contradicting $ad - bc = 1$; and the sign $+$ forces $c = 0$ and $a = d$. So $M = \pm T^m$ for some m. Then $ST^mS^{-1} = \left(\begin{smallmatrix} 1 & 0 \\ -m & 1 \end{smallmatrix}\right)$ must equal $\pm T^m$, which forces $m = 0$. So $M = \pm I$, and the center of $PSL_2(\mathbb{Z})$ is trivial. Hence $Z(B_3) \subseteq K$. □

Exercise 12.4.6 (Generators, presentations, and Cayley graphs)

(1) Draw the Cayley graph of S_3 with respect to the generators $a = (1\,2\,3)$ and $b = (1\,2)$, with undirected b-edges since $b = b^{-1}$. Find a closed path through e that expresses the relation $bab^{-1} = a^{-1}$.

(2) Show that $B_2 \cong \mathbb{Z}$, and that B_n is not abelian for $n \geqslant 3$.

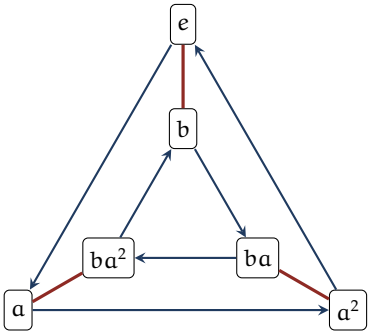
(3) Compute S^2, S^4, R^2, R^3, and R^6 in $SL_2(\mathbb{Z})$, where $R = ST$. Deduce that S and R have orders 4 and 6 in $SL_2(\mathbb{Z})$, and orders 2 and 3 in $PSL_2(\mathbb{Z})$.

(4) Express $M = \begin{pmatrix} 3 & 5 \\ 1 & 2 \end{pmatrix}$ as a word in S and T, following the proof of Proposition 12.4.2.

(5) Let $\sigma \in S_4$ have one-line notation 2431. Find $\mathrm{inv}(\sigma)$, all the shortest words for σ in s_1, s_2, s_3, and the normal form $N(\sigma)$.

Solution. **(1)** The elements are e, $a = (1\,2\,3)$, $a^2 = (1\,3\,2)$, $b = (1\,2)$, $ba = (2\,3)$, and $ba^2 = (1\,3)$. The a-edges $g \to ga$ form two directed triangles, $e \to a \to a^2 \to e$ and $b \to ba \to ba^2 \to b$. The b-edges join g to

gb: since $ab = (1\,3) = ba^2$ and $a^2b = (2\,3) = ba$, they join e to b, a to ba^2, and a^2 to ba. So the Cayley graph is a triangular prism whose two triangles are traversed in opposite directions:



Here the a-edges are blue arrows and the b-edges are red. Starting at e and following the edges labeled b, a, b, a visits $e, b, ba, bab = a^2, a^3 = e$: a closed path around one of the square faces. It expresses $baba = e$, that is, $bab^{-1} = a^{-1}$ (as $b^{-1} = b$). The two triangles express $a^3 = e$.

(2) For $n = 2$ there is one generator s_1 and no relations (there are no indices with $|i - j| \geqslant 2$, and no braid relations, which need $i \leqslant n - 2 = 0$). So $B_2 = \langle s_1 \mid \rangle$ is the free group on one generator, which is isomorphic to $\mathbb{Z}$ (Section 3.6). For $n \geqslant 3$, the surjection $B_n \to S_n$ sends s_1s_2 and s_2s_1 to $(1\,2\,3)$ and $(1\,3\,2)$, which are different; so $s_1s_2 \neq s_2s_1$ in B_n.

(3) $S^2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^2 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -I$, so $S^4 = I$. With $R = ST = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}$,

$$R^2 = \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix}, \qquad R^3 = R \cdot R^2 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -I, \qquad R^6 = I.$$

Since $S, S^2, S^3 = -S$ are not I, S has order 4; since $R, R^2, R^3 = -I, R^4 = -R, R^5 = -R^2$ are not I, R has order 6. In $PSL_2(\mathbb{Z})$, $S^2 = R^3 = -I$ is the identity while S, R, R^2 are not $\pm I$, so the orders are 2 and 3.

(4) Here $c = 1 \neq 0$. Dividing, $a = 3 = 3 \cdot 1 + 0$, so $q = 3$ and $r = 0$:

$$T^{-3}M = \begin{pmatrix} 1 & -3 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 3 & 5 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 2 \end{pmatrix},$$

$$S^{-1}T^{-3}M = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} = T^2.$$

So $M = T^3ST^2$. Check:

$$T^3ST^2 = \begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 5 \\ 1 & 2 \end{pmatrix}.$$

(5) The inversions of $\sigma = 2431$ are the pairs of positions $(1,4)$, $(2,3)$, $(2,4)$, $(3,4)$, so $\mathrm{inv}(\sigma) = \ell(\sigma) = 4$. We use Proposition 12.2.5 repeatedly, writing each permutation in one-line notation and recording its ending set:

- 2431 has ending set $\{2,3\}$;
- $2431 \cdot s_2 = 2341$ has ending set $\{3\}$; then 2314 has $\{2\}$, then 2134 has $\{1\}$, then 1234;
- $2431 \cdot s_3 = 2413$ has ending set $\{2\}$; then 2143 has $\{1,3\}$; both $2143 \cdot s_1 = 1243$ and $2143 \cdot s_3 = 2134$ are single generators (s_3 and s_1).

Reading the words from the right, the shortest words for σ are

$$s_1 s_2 s_3 s_2, \qquad s_3 s_1 s_2 s_3, \qquad s_1 s_3 s_2 s_3.$$

They are related by the relations of Theorem 3.6.18: $s_2 s_3 s_2 = s_3 s_2 s_3$ and $s_1 s_3 = s_3 s_1$. The normal form takes the smallest index at each step, $2431 \to 2341 \to 2314 \to 2134 \to 1234$ via s_2, s_3, s_2, s_1, so $N(\sigma) = s_1 s_2 s_3 s_2$. (In cycle notation $\sigma = (1\,2\,4)$, an even permutation, as $(-1)^4 = 1$ predicts.)

PART IV

Representation Theory

CHAPTER 13

Representations and Characters

Representation theory studies a group through its actions on vector spaces, that is, through homomorphisms $G \to \mathrm{GL}(V)$. This turns questions about an abstract group into linear algebra, where eigenvalues, traces and diagonalization are available. Historically the order was the reverse: groups first arose as groups of symmetries, and in the nineteenth century they were thought of as subgroups of GL_n from the start. The modern point of view separates the study of abstract groups, which occupied the earlier chapters, from the study of the ways in which an abstract group can be realized by matrices, which begins here.

In this chapter and the next we work over $k = \mathbb{C}$, and all representations are on finite-dimensional vector spaces. Several arguments use eigenvalues or complex conjugation, so the choice of $\mathbb{C}$ matters. From Section 13.2 on, G is a finite group.

13.1 Representations and Subrepresentations

Definition 13.1.1 (Representation)

A ***representation*** of a group G is a vector space V together with a group homomorphism $\rho : G \to \mathrm{GL}(V)$. Equivalently, it is an action of G on V in which every $g \in G$ acts by a linear map $V \to V$. We write gv or $g \cdot v$ for $\rho(g)(v)$, and we often call V itself the representation. The ***dimension*** of the representation is $\dim V$.

Definition 13.1.2 (Subrepresentation and irreducibility)

Let V be a representation of G.

- A ***subrepresentation*** of V is a subspace $W \subseteq V$ that is G-invariant: $gW = W$ for all $g \in G$. (It suffices that $gW \subseteq W$ for all g, since then also

$g^{-1}W \subseteq W$, that is, $W \subseteq gW$.) A subrepresentation is a representation of G in its own right.
- V is ***irreducible*** if $V \neq 0$ and the only subrepresentations of V are $\{0\}$ and V.

Definition 13.1.3 (Homomorphism of representations)

Let V and W be representations of G. A ***homomorphism of representations*** (or G-***equivariant map***) is a linear map $\varphi : V \to W$ with $\varphi(gv) = g\varphi(v)$ for all $g \in G$ and $v \in V$; equivalently, $g \circ \varphi = \varphi \circ g$ for every g, where g stands for the operator by which g acts on V or on W. The space of such maps is denoted $\mathrm{Hom}_G(V, W)$. An ***isomorphism of representations*** is a bijective homomorphism of representations; its inverse is again equivariant. We write $V \cong W$ if there is one.

NOTE If $\varphi \in \mathrm{Hom}_G(V, W)$, then $\ker \varphi$ and $\mathrm{Im}\, \varphi$ are subrepresentations of V and W. For instance, if $v \in \ker \varphi$, then $\varphi(gv) = g\varphi(v) = 0$, so $gv \in \ker \varphi$; and $g\varphi(v) = \varphi(gv)$ lies in $\mathrm{Im}\, \varphi$.

A 1-dimensional representation of a group G is a homomorphism $\theta : G \to \mathrm{GL}_1(\mathbb{C}) = \mathbb{C}^*$. Two of them, θ and θ', are isomorphic only if they are equal: an isomorphism $\mathbb{C} \to \mathbb{C}$ is multiplication by some $c \in \mathbb{C}^*$, and $c\,\theta(g) = \theta'(g)\,c$ forces $\theta(g) = \theta'(g)$.

Example 13.1.4 (First examples)

- The ***trivial representation*** of G is $\mathbb{C}$ with every g acting as the identity.
- If G acts on a finite set S, the ***permutation representation*** $\mathbb{C}[S]$ has basis $\{e_s\}_{s \in S}$, and g acts by $ge_s = e_{gs}$. Each g acts by a permutation matrix.
- Let $G = S_n$. Besides the trivial representation, which we denote U, there is the ***alternating*** (or ***sign***) representation $U' = \mathbb{C}$, on which σ acts as multiplication by $\mathrm{sgn}(\sigma)$ (Definition 3.4.10); it is a representation because sgn is a homomorphism. The permutation representation of S_n on $\{1, \dots, n\}$ is $\mathbb{C}^n$ with $\sigma e_i = e_{\sigma(i)}$. The line spanned by $e_1 + \cdots + e_n$ is a subrepresentation isomorphic to U. The subspace

$$V = \{(z_1, \dots, z_n) \in \mathbb{C}^n : z_1 + \cdots + z_n = 0\}$$

is also a subrepresentation, because permuting coordinates does not change their sum. Since $\mathbb{C}^n = \mathbb{C}(e_1 + \cdots + e_n) \oplus V$, the subrepresentation V is a complement of dimension $n-1$. It is called the ***standard representation*** of S_n, and it is irreducible: we prove this for $n = 3$ in Section 13.3.1, and in general in the exercises at the end of the chapter.

13.1.1 Representations of Finite Abelian Groups

The basic linear-algebra fact behind this subsection is that an operator of finite order can be diagonalized.

Lemma 13.1.5 (Operators of finite order)

Let V be a finite-dimensional complex vector space and $\varphi \in \mathrm{End}(V)$ an operator with $\varphi^n = \mathrm{id}_V$ for some $n \geqslant 1$. Then φ is diagonalizable, and its eigenvalues are nth roots of unity.

Proof. If $\varphi(v) = \lambda v$ with $v \neq 0$, then $v = \varphi^n(v) = \lambda^n v$, so $\lambda^n = 1$. By Theorem 7.6.11, $V = \bigoplus_\lambda V_\lambda$, where λ runs over the eigenvalues of φ and $V_\lambda = \ker(\varphi - \lambda)^N$ with $N = \dim V$. Fix an eigenvalue λ. For $\mu \neq \lambda$, the restriction of $\varphi - \mu$ to V_λ is injective: if $v \in V_\lambda$ and $\varphi(v) = \mu v$, then $0 = (\varphi-\lambda)^N v = (\mu-\lambda)^N v$, so $v = 0$.

Let $\lambda_k = e^{2\pi i k/n}$ for $0 \leqslant k \leqslant n-1$, so that $x^n - 1 = \prod_{k=0}^{n-1}(x - \lambda_k)$ and therefore $0 = \varphi^n - \mathrm{id}_V = \prod_{k=0}^{n-1}(\varphi - \lambda_k)$. The eigenvalue λ is one of the λ_k; write the product as $P \circ (\varphi - \lambda)$, where P is the product of the other factors. All the factors commute with φ, so they preserve V_λ, and P is injective on V_λ. Hence $(\varphi - \lambda)v = 0$ for every $v \in V_\lambda$: the space V_λ is the λ-eigenspace, and V is the direct sum of the eigenspaces of φ. □

Example 13.1.6 (Representations of cyclic groups)

Let $G = \mathbb{Z}/n$. A representation of G amounts to a vector space V with an operator $\varphi = \rho(1)$ such that $\varphi^n = \mathrm{id}_V$, since $\rho(m) = \varphi^m$. By Lemma 13.1.5, V has a basis $v_1, \ldots, v_d$ of eigenvectors of φ. Each line $\mathbb{C}v_j$ is invariant under every $\rho(m) = \varphi^m$, so it is a 1-dimensional, hence irreducible, subrepresentation, and $V = \mathbb{C}v_1 \oplus \cdots \oplus \mathbb{C}v_d$.

A 1-dimensional representation of $\mathbb{Z}/n$ is a homomorphism $\mathbb{Z}/n \to \mathrm{GL}_1(\mathbb{C}) = \mathbb{C}^*$. It is determined by the image of 1, which must be an nth root of unity $e^{2\pi i k/n}$, and each choice of $k \in \{0, \ldots, n-1\}$ gives one. These n representations are pairwise non-isomorphic, since 1-dimensional representations are isomorphic only if they are equal. So $\mathbb{Z}/n$ has exactly n

irreducible representations up to isomorphism, and every representation of $\mathbb{Z}/n$ is a direct sum of copies of them.

Proposition 13.1.7 (Representations of finite abelian groups)

Let G be a finite abelian group and V a representation of G. Then V has a basis consisting of vectors that are eigenvectors of every $g \in G$. Consequently V is a direct sum of 1-dimensional subrepresentations, and every irreducible representation of G is 1-dimensional.

Proof. We show by induction on m that commuting operators $\psi_1, \dots, \psi_m$ of finite order on a finite-dimensional complex vector space V have a common basis of eigenvectors. For $m = 1$ this is Lemma 13.1.5. For $m > 1$, write $V = \bigoplus_\lambda E_\lambda$ as the direct sum of the eigenspaces of ψ_1. Each E_λ is invariant under every ψ_j: if $v \in E_\lambda$, then $\psi_1(\psi_j v) = \psi_j(\psi_1 v) = \lambda \psi_j v$. The restrictions of $\psi_2, \dots, \psi_m$ to E_λ commute and have finite order, so by induction E_λ has a basis of common eigenvectors of $\psi_2, \dots, \psi_m$. These are also eigenvectors of ψ_1, which acts on E_λ as the scalar λ. Putting these bases together gives a basis of V of common eigenvectors.

Apply this to the operators $\rho(g)$, $g \in G$, which commute because G is abelian and have finite order because G is finite. The line spanned by each common eigenvector is invariant under every g, hence a 1-dimensional subrepresentation, and V is the direct sum of these lines. If V is irreducible, it must be one of them. □

Definition 13.1.8 (Dual group)

For a finite abelian group G, the ***dual group*** is $\widehat{G} = \mathrm{Hom}(G, \mathbb{C}^*)$, the set of homomorphisms $G \to \mathbb{C}^*$, with pointwise multiplication: $(\theta\theta')(g) = \theta(g)\theta'(g)$.

Because $\mathbb{C}^*$ is abelian, $\theta\theta'$ is again a homomorphism:

$$(\theta\theta')(gh) = \theta(g)\theta(h)\theta'(g)\theta'(h) = (\theta\theta')(g)\,(\theta\theta')(h).$$

So $\widehat{G}$ is an abelian group, with identity the constant function 1 and inverses $\theta^{-1}(g) = \theta(g)^{-1}$. By Proposition 13.1.7, the irreducible representations of a finite abelian group G are, up to isomorphism, exactly the elements of $\widehat{G}$, and every representation of G is a direct sum of them. This completes the classification of the complex representations of finite abelian groups.

Proposition 13.1.9 (The dual of a finite abelian group)
If G is a finite abelian group, then $\widehat{G} \cong G$.

Proof. By Theorem 10.2.1, $G \cong \mathbb{Z}/m_1 \times \cdots \times \mathbb{Z}/m_r$; let $\varepsilon_1, \ldots, \varepsilon_r$ be the standard generators. A homomorphism $\theta : G \to \mathbb{C}^*$ is determined by the values $\zeta_i = \theta(\varepsilon_i)$, which satisfy $\zeta_i^{m_i} = \theta(m_i \varepsilon_i) = 1$. Conversely, for any m_ith roots of unity $\zeta_1, \ldots, \zeta_r$, the formula $\theta(a_1, \ldots, a_r) = \zeta_1^{a_1} \cdots \zeta_r^{a_r}$ is well defined (each $\zeta_i^{a_i}$ depends only on a_i modulo m_i) and is a homomorphism. Writing μ_m for the group of mth roots of unity, $\theta \mapsto (\theta(\varepsilon_1), \ldots, \theta(\varepsilon_r))$ is therefore an isomorphism $\widehat{G} \cong \mu_{m_1} \times \cdots \times \mu_{m_r}$, and $\mu_m \cong \mathbb{Z}/m$ via $e^{2\pi i k/m} \leftrightarrow k$. □

NOTE The isomorphism $\widehat{G} \cong G$ depends on a choice of generators. There is no canonical isomorphism, even for $G = \mathbb{Z}/n$.

13.1.2 Constructions with Representations

For a representation X of G, the subspace of ***invariant vectors*** is $X^G = \{x \in X : gx = x \text{ for all } g \in G\}$.

Proposition 13.1.10 (Constructions with representations)
Let V and W be representations of G. Each of the following is a representation of G.

(i) The direct sum $V \oplus W$, with $g(v, w) = (gv, gw)$.
(ii) The tensor product $V \otimes W$ (Definition 9.8.5), with $g(v \otimes w) = gv \otimes gw$, extended linearly.
(iii) The ***dual representation*** $V^* = \mathrm{Hom}(V, \mathbb{C})$, with $(g\ell)(v) = \ell(g^{-1}v)$. Thus g acts on V^* by the dual map $\rho(g^{-1})^t$ (Definition 6.2.9).
(iv) $\mathrm{Hom}(V, W)$, with $g \cdot \varphi = g \circ \varphi \circ g^{-1}$ (the first g acting on W, and g^{-1} on V). Its invariant vectors are exactly the homomorphisms of representations: $\mathrm{Hom}(V, W)^G = \mathrm{Hom}_G(V, W)$.
(v) For $d \geqslant 1$, the symmetric and exterior powers $\mathrm{Sym}^d V$ and $\bigwedge^d V$ (Definitions 9.10.3 and 9.10.4), with $g(v_1 \cdots v_d) = (gv_1) \cdots (gv_d)$ and $g(v_1 \wedge \cdots \wedge v_d) = gv_1 \wedge \cdots \wedge gv_d$.
(vi) If $W \subseteq V$ is a subrepresentation, the quotient V/W, with $g(v + W) = gv + W$.

Moreover, the canonical isomorphism $V^* \otimes W \to \mathrm{Hom}(V, W)$, $\ell \otimes w \mapsto (v \mapsto \ell(v)w)$, of Proposition 9.8.10 is an isomorphism of representations.

Proof. In each case g acts by a linear map: in (ii) by $\rho_V(g) \otimes \rho_W(g)$ (Proposition 9.8.7); in (v) by $\mathrm{Sym}^d \rho(g)$ and $\bigwedge^d \rho(g)$ (Proposition 9.10.6); in (vi) because $gW \subseteq W$. The identity acts as the identity, and it remains to check that gh acts as g after h. In (i), (ii), (v) and (vi) this is clear on the vectors written, which span. In (iii), $(g(h\ell))(v) = (h\ell)(g^{-1}v) = \ell(h^{-1}g^{-1}v) = ((gh)\ell)(v)$. In (iv), $g \cdot (h \cdot \varphi) = gh\varphi h^{-1}g^{-1} = (gh) \cdot \varphi$. Also $g\varphi g^{-1} = \varphi$ for all g if and only if $g\varphi = \varphi g$ for all g, which proves $\mathrm{Hom}(V,W)^G = \mathrm{Hom}_G(V,W)$.

Finally, $g(\ell \otimes w) = (\ell \circ g^{-1}) \otimes gw$ is sent to the map $v \mapsto \ell(g^{-1}v)\, gw$, which is $g \circ (v \mapsto \ell(v)w) \circ g^{-1}$. So the isomorphism commutes with the action of every g. □

13.2 Maschke's Theorem and Complete Reducibility

From now on G is a finite group. The basic tool is averaging over G, which makes sense because G is finite.

Lemma 13.2.1 (Invariant inner products)

Let V be a representation of G. There is a positive-definite Hermitian inner product H on V (Definition 9.1.16) that is G-invariant: $H(gv, gw) = H(v, w)$ for all $g \in G$ and $v, w \in V$. In other words, every g acts on V by a unitary operator with respect to H.

Proof. Let H_0 be any positive-definite Hermitian inner product on V; for instance, choose a basis $e_1, \dots, e_n$ and set $H_0(\sum a_i e_i, \sum b_i e_i) = \sum_i \overline{a_i} b_i$. Average it over G:

$$H(v,w) = \frac{1}{|G|} \sum_{g \in G} H_0(gv, gw).$$

Each term is conjugate-linear in v, linear in w, and satisfies $H_0(gw, gv) = \overline{H_0(gv, gw)}$, so H is a Hermitian form. It is positive definite: for $v \neq 0$, every term $H_0(gv, gv)$ is positive, since $gv \neq 0$. It is G-invariant: for $h \in G$,

$$H(hv, hw) = \frac{1}{|G|} \sum_{g \in G} H_0(ghv, ghw) = \frac{1}{|G|} \sum_{g' \in G} H_0(g'v, g'w) = H(v,w),$$

because $g' = gh$ runs over G exactly once as g does. □

Theorem 13.2.2 (Maschke's Theorem)

Let V be a representation of a finite group G, and let $W \subseteq V$ be a subrepresentation. Then there is a subrepresentation $U \subseteq V$ with $V = W \oplus U$.

Proof. Equip V with a G-invariant Hermitian inner product H (Lemma 13.2.1), and let $U = W^{\perp} = \{v \in V : H(w, v) = 0 \text{ for all } w \in W\}$. By Theorem 9.1.31, $V = W \oplus W^{\perp}$. If $v \in W^{\perp}$, $g \in G$ and $w \in W$, then $g^{-1}w \in W$, so

$$H(w, gv) = H(gg^{-1}w, gv) = H(g^{-1}w, v) = 0.$$

Hence $gv \in W^{\perp}$, and $U = W^{\perp}$ is a subrepresentation. □

There is a second proof, which uses no inner product. It works over any field k in which $|G| \neq 0$, that is, with $\operatorname{char} k \nmid |G|$.

Second proof. Choose any complementary subspace U_0, so that $V = W \oplus U_0$, and let $\pi_0 : V \to W$ be the projection onto W along U_0 (so $\pi_0|_W = \mathrm{id}_W$ and $\pi_0|_{U_0} = 0$). Average it:

$$\pi(v) = \frac{1}{|G|} \sum_{g \in G} g\, \pi_0(g^{-1}v).$$

Since π_0 takes values in W and $gW = W$, so does π. For $w \in W$, each $g^{-1}w$ lies in W, so $g\pi_0(g^{-1}w) = gg^{-1}w = w$, and $\pi(w) = w$. The map π is G-equivariant: for $h \in G$, substituting $g = hg'$,

$$\pi(hv) = \frac{1}{|G|} \sum_{g \in G} g\, \pi_0(g^{-1}hv) = \frac{1}{|G|} \sum_{g' \in G} hg'\, \pi_0(g'^{-1}v) = h\, \pi(v).$$

Let $U = \ker \pi$, a subrepresentation because π is equivariant. Every $v \in V$ is $\pi(v) + (v - \pi(v))$ with $\pi(v) \in W$ and $\pi(v - \pi(v)) = \pi(v) - \pi(v) = 0$, because π is the identity on W. Also $W \cap U = 0$, again because π is the identity on W. So $V = W \oplus U$. □

NOTE Both hypotheses matter. If $G = \mathbb{Z}$ acts on $\mathbb{C}^2$ by $t \mapsto \left(\begin{smallmatrix} 1 & t \\ 0 & 1 \end{smallmatrix}\right)$, the line $\mathbb{C} \times \{0\}$ is invariant, but it is the only invariant line (it is the only line of eigenvectors of $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$), so it has no invariant complement; the same happens for $\mathbb{R}$ acting by the same matrices. Here averaging is impossible because the group is infinite. Similarly, if k has characteristic p, the same matrices define a representation of $\mathbb{Z}/p$ on k^2 with no invariant complement to $k \times \{0\}$. This is one reason why *modular representation theory*, over fields whose characteristic divides $|G|$, is harder. On the other hand, averaging still works for compact groups such as the circle S^1 or $SO(n)$, with the sum over G replaced by an integral, but we will not pursue this.

Corollary 13.2.3 (Complete reducibility)

Every representation of a finite group G is a direct sum of irreducible subrepresentations.

Proof. Induct on $\dim V$. If $V = 0$, it is the empty direct sum, and if V is irreducible there is nothing to prove. Otherwise V has a subrepresentation W with $0 \neq W \neq V$, and by Maschke's Theorem $V = W \oplus U$ for a subrepresentation U. Both W and U have smaller dimension than V, so each is a direct sum of irreducible subrepresentations, and hence so is V. □

13.3 Schur's Lemma

Our goal is to find the irreducible representations of a given group G and to describe how every representation decomposes into them. The key tool is the following lemma.

Lemma 13.3.1 (Schur's Lemma)

Let V and W be irreducible representations of G, and let $\varphi : V \to W$ be a homomorphism of representations.

(i) Either $\varphi = 0$ or φ is an isomorphism.

(ii) If $W = V$, then $\varphi = \lambda \operatorname{id}_V$ for some $\lambda \in \mathbb{C}$.

Proof. (i) The kernel $\ker \varphi$ is a subrepresentation of V, so either $\ker \varphi = V$, in which case $\varphi = 0$, or $\ker \varphi = 0$ and φ is injective. Similarly $\operatorname{Im} \varphi$ is a subrepresentation of W, so either $\operatorname{Im} \varphi = 0$, in which case $\varphi = 0$, or $\operatorname{Im} \varphi = W$ and φ is surjective. So if $\varphi \neq 0$, it is bijective.

(ii) Since V is a nonzero finite-dimensional complex vector space, φ has an eigenvalue λ (Theorem 7.5.1). Then $\varphi - \lambda \operatorname{id}_V$ is again a homomorphism of representations $V \to V$, and it has nonzero kernel, so it is not an isomorphism. By (i), $\varphi - \lambda \operatorname{id}_V = 0$. □

Corollary 13.3.2 (Homomorphisms between irreducibles)

If V and W are irreducible representations of G, then

$$\dim \operatorname{Hom}_G(V, W) = \begin{cases} 1 & \text{if } V \cong W, \\ 0 & \text{if } V \ncong W. \end{cases}$$

Proof. If $V \ncong W$, every homomorphism $V \to W$ is zero by Schur's Lemma. If $\psi : V \to W$ is an isomorphism of representations, then $\varphi \mapsto \psi^{-1} \circ \varphi$ is a

linear isomorphism $\mathrm{Hom}_G(V,W) \to \mathrm{Hom}_G(V,V)$, and $\mathrm{Hom}_G(V,V) = \mathbb{C}\,\mathrm{id}_V$ by Schur's Lemma. □

Example 13.3.3 (The center acts by scalars)

Let V be an irreducible representation of G and $h \in Z(G)$, the center of G (Definition 3.1.13). Since h commutes with every $g \in G$, the operator $\rho(h)$ commutes with every $\rho(g)$, that is, $\rho(h) \in \mathrm{Hom}_G(V,V)$. By Schur's Lemma, $\rho(h) = \lambda_h\,\mathrm{id}_V$ for some $\lambda_h \in \mathbb{C}$: every central element acts on every irreducible representation by a scalar. In particular, if G is abelian, every element acts on an irreducible V by a scalar, so every line in V is a subrepresentation and $\dim V = 1$. This is a second proof of the last statement of Proposition 13.1.7.

13.3.1 The Irreducible Representations of S_3

We now find all the irreducible representations of S_3, the smallest nonabelian group. We know three of them from Example 13.1.4: the trivial representation U, the alternating representation U', and the 2-dimensional standard representation $V \subset \mathbb{C}^3$. What is special about S_3 is that there are no others. Throughout, $\tau = (123)$, $\sigma = (12)$ and $\lambda = e^{2\pi i/3}$, so that $\lambda^3 = 1$, $\lambda^{-1} = \lambda^2$ and $1 + \lambda + \lambda^2 = 0$.

Two facts about S_3 do all the work. First, every element of S_3 is τ^j or $\sigma\tau^j$ for some $j \in \{0,1,2\}$ (these six elements are distinct). So a subspace of a representation that is invariant under σ and τ is a subrepresentation, and a linear map between representations that commutes with σ and τ is a homomorphism of representations. Second, conjugating a cycle by σ applies σ to its entries (Lemma 3.4.5), so $\sigma\tau\sigma^{-1} = (213) = \tau^{-1}$; since $\sigma = \sigma^{-1}$, this gives $\tau\sigma = \sigma\tau^{-1}$. Consequently, in any representation of S_3,

$$\tau v = \mu v \implies \tau(\sigma v) = \sigma(\tau^{-1} v) = \mu^{-1}\,\sigma v. \tag{$*$}$$

That is, σ maps the μ-eigenspace of τ into the μ^{-1}-eigenspace. By Lemma 13.1.5, τ is diagonalizable in every representation of S_3, with eigenvalues among $1, \lambda, \lambda^2$.

Lemma 13.3.4 (The standard representation of S_3)

Let $u = e_1 + \lambda^2 e_2 + \lambda e_3$. Then u and $\sigma u = \lambda^2 e_1 + e_2 + \lambda e_3$ form a basis of V, and

$$\tau u = \lambda u, \qquad \tau(\sigma u) = \lambda^2\,\sigma u, \qquad \sigma(\sigma u) = u.$$

So τ acts on V with eigenvalues λ and λ^2, and σ with eigenvalues 1 and -1 (with eigenvectors $u + \sigma u$ and $u - \sigma u$). Moreover, V is irreducible.

Proof. Both vectors lie in V because $1+\lambda+\lambda^2=0$. Since $\tau e_1=e_2$, $\tau e_2=e_3$ and $\tau e_3=e_1$, we get $\tau u=\lambda e_1+e_2+\lambda^2 e_3=\lambda u$, using $\lambda^3=1$. The transposition σ swaps e_1 and e_2, which gives the formula for σu, and $\sigma(\sigma u)=\sigma^2 u=u$. By $(*)$, σu is an eigenvector of τ for the eigenvalue $\lambda^{-1}=\lambda^2$. Eigenvectors for distinct eigenvalues are linearly independent, and $\dim V=2$, so u, σu is a basis of V.

Let $X\subseteq V$ be a nonzero subrepresentation. The restriction of τ to X is diagonalizable (Lemma 13.1.5), so X contains an eigenvector of τ. The eigenspaces of τ on V are the lines $\mathbb{C}u$ and $\mathbb{C}\sigma u$, so X contains u or σu. Being σ-invariant, it contains both, and $X=V$. □

Proposition 13.3.5 (The irreducible representations of S_3)

Every irreducible representation of S_3 is isomorphic to exactly one of U, U', V. Every representation W of S_3 is isomorphic to $U^{\oplus a}\oplus U'^{\oplus b}\oplus V^{\oplus c}$ for unique integers $a,b,c\geqslant 0$, and these are determined by

$$a+b=\dim\ker(\tau-1),\qquad c=\dim\ker(\tau-\lambda),\qquad a+c=\dim\ker(\sigma-1),$$

where τ and σ denote the operators by which (123) and (12) act on W.

Proof. Let W be irreducible, and let $v\in W$ be an eigenvector of τ, say $\tau v=\mu v$. By $(*)$, σv is also an eigenvector of τ, so $\operatorname{span}(v,\sigma v)$ is invariant under τ; it is invariant under σ because $\sigma^2=e$. It is therefore a nonzero subrepresentation, and $W=\operatorname{span}(v,\sigma v)$. In particular, every irreducible representation of S_3 has dimension at most 2.

Case $\mu=1$. Then $\tau(\sigma v)=\sigma v$ as well. If v and σv were linearly independent, then $w=v+\sigma v$ would be a nonzero vector with $\sigma w=w$ and $\tau w=w$, so $\mathbb{C}w$ would be a subrepresentation of the 2-dimensional space W, contradicting irreducibility. So $\sigma v=cv$ for some $c\in\mathbb{C}$, and $c^2=1$ because $\sigma^2=e$. Then $W=\mathbb{C}v$, with τ acting as 1 and σ as $c=\pm1$. Since $\operatorname{sgn}\tau=1$ and $\operatorname{sgn}\sigma=-1$, the linear map $v\mapsto 1$ is an isomorphism $W\cong U$ if $c=1$ and $W\cong U'$ if $c=-1$.

Case $\mu\neq 1$. Then $\mu\in\{\lambda,\lambda^2\}$. If $\mu=\lambda^2$, replace v by σv, which is a λ-eigenvector by $(*)$ and satisfies $\sigma(\sigma v)=v$; so we may assume $\mu=\lambda$. Then v and σv are eigenvectors of τ for the distinct eigenvalues λ and λ^2, hence a basis of W. Define a linear isomorphism $\psi: W\to V$ by $\psi(v)=u$ and $\psi(\sigma v)=\sigma u$, with u as in Lemma 13.3.4. Both bases are eigenbases of τ with the same eigenvalues, and σ swaps the two vectors in each basis, so ψ commutes with τ and σ. Hence ψ is an isomorphism of representations, and $W\cong V$.

The representations U, U', V are irreducible (Lemma 13.3.4 for V) and pairwise non-isomorphic: V has dimension 2, and σ acts as 1 on U and as -1 on U'. This proves the first statement.

Now let W be any representation. By complete reducibility and the first statement, $W \cong U^{\oplus a} \oplus U'^{\oplus b} \oplus V^{\oplus c}$ for some $a, b, c \geqslant 0$. The operator τ acts as 1 on U and on U', and on V with eigenvalues λ and λ^2, each with a 1-dimensional eigenspace. The operator σ acts as 1 on U, as -1 on U', and on V with eigenvalues 1 and -1, each with a 1-dimensional eigenspace. Adding up the dimensions of the eigenspaces summand by summand gives the three formulas. Their left-hand sides depend only on the isomorphism class of W, and they determine c, then a, then b; so a, b, c are unique. □

Example 13.3.6 (Decomposing $V \otimes V$, $\operatorname{Sym}^2 V$ and $\bigwedge^2 V$)

Let $u_1 = u$ and $u_2 = \sigma u$ be the basis of V from Lemma 13.3.4, so that $\tau u_1 = \lambda u_1$, $\tau u_2 = \lambda^2 u_2$, and σ swaps u_1 and u_2. The basis

$$u_1 \otimes u_1, \quad u_1 \otimes u_2, \quad u_2 \otimes u_1, \quad u_2 \otimes u_2$$

of $V \otimes V$ consists of eigenvectors of τ, with eigenvalues $\lambda^2, 1, 1, \lambda$ (recall $\lambda^4 = \lambda$). So $a + b = 2$ and $c = 1$ in Proposition 13.3.5. The operator σ swaps $u_1 \otimes u_1$ with $u_2 \otimes u_2$, and $u_1 \otimes u_2$ with $u_2 \otimes u_1$, so it has eigenvalues $1, 1, -1, -1$, and $a + c = 2$. Hence $a = b = c = 1$:

$$V \otimes V \cong U \oplus U' \oplus V.$$

Explicitly, $u_1 \otimes u_2 + u_2 \otimes u_1$ spans a copy of U, and $u_1 \otimes u_2 - u_2 \otimes u_1$ spans a copy of U'.
Similarly, $\operatorname{Sym}^2 V$ has basis u_1^2, u_1u_2, u_2^2, on which τ has eigenvalues $\lambda^2, 1, \lambda$, while σ fixes u_1u_2 and swaps u_1^2 and u_2^2. So $a + b = 1$, $c = 1$ and $a + c = 2$, and $\operatorname{Sym}^2 V \cong U \oplus V$. Finally, $\bigwedge^2 V$ is spanned by $u_1 \wedge u_2$, which τ multiplies by $\lambda \cdot \lambda^2 = 1$ and σ sends to $u_2 \wedge u_1 = -u_1 \wedge u_2$; so $\bigwedge^2 V \cong U'$. This is to be expected: on the top exterior power $\bigwedge^2 V$, each g acts by $\det(g|_V)$ (Theorem 9.11.3), and $\det(\sigma|_V) = 1 \cdot (-1) = -1 = \operatorname{sgn}\sigma$, $\det(\tau|_V) = \lambda \cdot \lambda^2 = 1 = \operatorname{sgn}\tau$.

13.4 Characters

For S_3 we decomposed representations by looking at the eigenvalues of two group elements. For a general group this is a lot of information to organize, and it turns out that much less suffices: the traces of the group elements.

Definition 13.4.1 (Character)

The ***character*** of a representation V of G is the function $\chi_V : G \to \mathbb{C}$ given by $\chi_V(g) = \mathrm{Tr}(\rho(g))$, the trace (Definition 5.3.11) of the operator by which g acts on V.

Definition 13.4.2 (Class function)

A ***class function*** on G is a function $f : G \to \mathbb{C}$ with $f(hgh^{-1}) = f(g)$ for all $g, h \in G$, that is, a function that is constant on each conjugacy class of G (Definition 3.2.12).

Since $\mathrm{Tr}(ABA^{-1}) = \mathrm{Tr}(B)$, we have $\chi_V(hgh^{-1}) = \mathrm{Tr}(\rho(h)\rho(g)\rho(h)^{-1}) = \chi_V(g)$: every character is a class function. Isomorphic representations have the same character, because an isomorphism $\psi : V \to W$ of representations satisfies $\rho_W(g) = \psi\,\rho_V(g)\,\psi^{-1}$.

Proposition 13.4.3 (Properties of characters)

Let V and W be representations of the finite group G, and let $g \in G$.

(i) $\chi_V(e) = \dim V$.

(ii) $\chi_{V\oplus W}(g) = \chi_V(g) + \chi_W(g)$ and $\chi_{V\otimes W}(g) = \chi_V(g)\,\chi_W(g)$.

(iii) $\chi_V(g^{-1}) = \overline{\chi_V(g)}$ and $\chi_{V^*}(g) = \overline{\chi_V(g)}$.

(iv) $\chi_{\mathrm{Hom}(V,W)}(g) = \overline{\chi_V(g)}\,\chi_W(g)$.

(v) If $V = \mathbb{C}[S]$ is the permutation representation of an action of G on a finite set S, then $\chi_V(g) = |S^g|$, the number of points of S fixed by g.

(vi) $\chi_{\mathrm{Sym}^2 V}(g) = \frac{1}{2}\big(\chi_V(g)^2 + \chi_V(g^2)\big)$ and $\chi_{\Lambda^2 V}(g) = \frac{1}{2}\big(\chi_V(g)^2 - \chi_V(g^2)\big)$.

Proof. Since g has finite order, Lemma 13.1.5 gives a basis $v_1, \dots, v_n$ of V with $gv_i = \lambda_i v_i$, where the λ_i are roots of unity; then $\chi_V(g) = \sum_i \lambda_i$. Similarly, let $w_1, \dots, w_m$ be a basis of W with $gw_j = \mu_j w_j$.

(i) The identity element acts as id_V, whose trace is $\dim V$.

(ii) In the basis $v_1, \dots, v_n, w_1, \dots, w_m$ of $V \oplus W$, the matrix of g is block diagonal, so the traces add. The vectors $v_i \otimes w_j$ form a basis of $V \otimes W$, and $g(v_i \otimes w_j) = \lambda_i\mu_j\,(v_i \otimes w_j)$, so $\chi_{V\otimes W}(g) = \sum_{i,j} \lambda_i\mu_j = \big(\sum_i \lambda_i\big)\big(\sum_j \mu_j\big)$.

(iii) We have $g^{-1}v_i = \lambda_i^{-1}v_i$, and $\lambda_i^{-1} = \overline{\lambda_i}$ because $|\lambda_i| = 1$; so $\chi_V(g^{-1}) = \sum_i \overline{\lambda_i} = \overline{\chi_V(g)}$. On V^*, g acts by the dual map $\rho(g^{-1})^t$, whose matrix in the dual basis is the transpose of the matrix of $\rho(g^{-1})$. A matrix and its transpose have the same trace, so $\chi_{V^*}(g) = \chi_V(g^{-1}) = \overline{\chi_V(g)}$.

(iv) By Proposition 13.1.10, $\mathrm{Hom}(V, W) \cong V^* \otimes W$ as representations, so this follows from (ii) and (iii).

(v) The matrix of g in the basis $\{e_s\}$ is a permutation matrix, and its diagonal entry in position s is 1 if $gs = s$ and 0 otherwise.

(vi) By Proposition 9.10.19, $\mathrm{Sym}^2 V$ has a basis $\{v_i v_j\}_{i \leqslant j}$ and $\bigwedge^2 V$ a basis $\{v_i \wedge v_j\}_{i<j}$, and g multiplies both $v_i v_j$ and $v_i \wedge v_j$ by $\lambda_i \lambda_j$. Since g^2 has eigenvalues λ_i^2,

$$\chi_V(g)^2 = \sum_i \lambda_i^2 + 2\sum_{i<j} \lambda_i \lambda_j = \chi_V(g^2) + 2\sum_{i<j} \lambda_i \lambda_j.$$

Hence $\chi_{\bigwedge^2 V}(g) = \sum_{i<j} \lambda_i \lambda_j = \frac{1}{2}\big(\chi_V(g)^2 - \chi_V(g^2)\big)$, and $\chi_{\mathrm{Sym}^2 V}(g) = \sum_i \lambda_i^2 + \sum_{i<j} \lambda_i \lambda_j = \frac{1}{2}\big(\chi_V(g)^2 + \chi_V(g^2)\big)$. □

NOTE For a 1-dimensional representation $\theta : G \to \mathbb{C}^*$, the character is θ itself, so it is multiplicative: $\chi(gh) = \chi(g)\chi(h)$. For higher-dimensional representations this fails; already $\chi_V(e) = \dim V$ differs from $\chi_V(e)^2$ when $\dim V > 1$. Characters turn the operations $\oplus$ and $\otimes$ on representations into sums and products of functions, but they do not respect the group law.

Definition 13.4.4 (Character table)

The ***character table*** of a finite group G lists the value of the character of each irreducible representation of G on each conjugacy class. Its rows are indexed by the irreducible representations, up to isomorphism, and its columns by the conjugacy classes, each labelled by a representative and the number of elements of the class.

Example 13.4.5 (The character table of S_3)

By Proposition 13.3.5, the irreducible representations of S_3 are U, U' and V. By Theorem 3.4.6, the conjugacy classes of S_3 are $\{e\}$, the three transpositions and the two 3-cycles.

	e	(12)	(123)
size	1	3	2
U	1	1	1
U'	1	-1	1
V	2	0	-1

The first column consists of the dimensions, and the rows of U and U' are the constant function 1 and sgn. For V, we can use the eigenvalues

from Lemma 13.3.4 (1 and -1 for (12); λ and λ^2, with $\lambda+\lambda^2=-1$, for (123)). Alternatively, $U \oplus V$ is the permutation representation $\mathbb{C}^3$, whose character counts fixed points and takes the values $(3,1,0)$ on the three classes; subtracting χ_U gives $\chi_V = (2,0,-1)$.
Characters give a fast way to decompose representations. The character of $V \otimes V$ is χ_V^2, with values $(4,0,1)$. If $V \otimes V \cong U^{\oplus a} \oplus U'^{\oplus b} \oplus V^{\oplus c}$, then $\chi_V^2 = a\chi_U + b\chi_{U'} + c\chi_V$, and evaluating on the three classes gives $a+b+2c=4$, $a-b=0$ and $a+b-c=1$. The only solution is $a=b=c=1$, as in Example 13.3.6.

13.4.1 Why Traces Suffice: Symmetric Polynomials

Here is a heuristic reason why the character should capture the eigenvalues. An unordered list of n complex numbers $\lambda_1,\dots,\lambda_n$, possibly with repetitions, can be recorded by the coefficients of the polynomial $\prod_{i=1}^{n}(x-\lambda_i)$, whose roots they are; by the fundamental theorem of algebra, every monic polynomial of degree n arises from exactly one such list. The coefficient of x^{n-k} in $\prod_i(x-\lambda_i)$ is $(-1)^k\sigma_k(\lambda_1,\dots,\lambda_n)$, where

$$\sigma_k(z_1,\dots,z_n) = \sum_{i_1<\cdots<i_k} z_{i_1}\cdots z_{i_k}, \qquad 1 \leqslant k \leqslant n,$$

are the ***elementary symmetric polynomials***. The group S_n acts on $\mathbb{C}[z_1,\dots,z_n]$ by permuting the variables, and a ***symmetric polynomial*** is a polynomial fixed by every element of S_n.

Theorem 13.4.6 (Fundamental theorem on symmetric polynomials)
The subring $\mathbb{C}[z_1,\dots,z_n]^{S_n}$ of symmetric polynomials is the polynomial ring $\mathbb{C}[\sigma_1,\dots,\sigma_n]$: every symmetric polynomial is a polynomial in $\sigma_1,\dots,\sigma_n$, in exactly one way.

We will not use this theorem, and we only indicate why it holds. For $n=2$, where $\sigma_1 = z_1+z_2$ and $\sigma_2 = z_1z_2$, the first few symmetric polynomials are

$$z_1^2+z_2^2 = \sigma_1^2-2\sigma_2, \qquad z_1^2z_2+z_1z_2^2 = \sigma_1\sigma_2, \qquad z_1^3+z_2^3 = \sigma_1^3-3\sigma_1\sigma_2.$$

In general, the nonconstant monomials of a symmetric polynomial $p(z_1,z_2)$ that are not divisible by z_1z_2 are powers of z_1 and of z_2, and they occur in pairs, so $p = a_0 + \sum_{k\geqslant 1} a_k(z_1^k+z_2^k) + z_1z_2\,q$. Since $(z_1+z_2)^k - z_1^k - z_2^k$ is divisible by z_1z_2 for $k \geqslant 1$, this can be rewritten as $p = a_0 + \sum_{k\geqslant 1} a_k\sigma_1^k + \sigma_2\,q'$, where q' is symmetric of lower degree than p; induction on the degree finishes the argument.

The theorem can also be viewed through representation theory. The homogeneous polynomials of degree 1 form the space $\operatorname{span}(z_1, \dots, z_n)$, on which S_n acts by its permutation representation $U \oplus V$, with invariant part the trivial summand spanned by σ_1. The homogeneous polynomials of degree d form Sym^d of this space (Theorem 9.10.15), and the symmetric ones among them are its invariant vectors, which make up the trivial summands in a decomposition of $\operatorname{Sym}^d(U \oplus V)$ into irreducibles. Carrying out a proof along these lines would require more about the representations of S_n than we develop.

Another family of symmetric polynomials is given by the power sums $p_k = z_1^k + \cdots + z_n^k$; for instance, $p_1 = \sigma_1$ and $p_2 = \sigma_1^2 - 2\sigma_2$. One can show in the same spirit that $p_1, \dots, p_n$ also generate: $\mathbb{C}[z_1, \dots, z_n]^{S_n} = \mathbb{C}[p_1, \dots, p_n]$. So an unordered list $\lambda_1, \dots, \lambda_n$ is determined by its power sums $\sum_i \lambda_i^k$, $1 \leqslant k \leqslant n$.

Now let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of g acting on V. Their power sums are $\sum_i \lambda_i^k = \operatorname{Tr}(\rho(g)^k) = \chi_V(g^k)$, and since G is a group, each g^k is again an element of G. So the single function χ_V determines the eigenvalues of every element of G on V. (There is also a direct argument: if g has order m, the multiplicity of a root of unity ζ as an eigenvalue of g is $\frac{1}{m}\sum_{j=0}^{m-1} \zeta^{-j}\chi_V(g^j)$, because $\sum_{j=0}^{m-1}(\lambda_i/\zeta)^j$ is m if $\lambda_i = \zeta$ and 0 otherwise.) We will see in the next section that the character determines much more: the representation itself, up to isomorphism.

13.5 Orthogonality of Characters

13.5.1 Invariants and the Inner Product on Class Functions

Proposition 13.5.1 (Averaging projection)
Let V be a representation of G, and let

$$\varphi_V = \frac{1}{|G|} \sum_{g \in G} g : V \to V.$$

Then φ_V is a projection onto V^G: its image is V^G, and it is the identity on V^G. Consequently,

$$\dim V^G = \operatorname{Tr}(\varphi_V) = \frac{1}{|G|} \sum_{g \in G} \chi_V(g).$$

Proof. If $v \in V^G$, every term gv equals v, so $\varphi_V(v) = v$. For any $v \in V$ and $h \in G$,

$$h\,\varphi_V(v) = \frac{1}{|G|} \sum_{g \in G} hgv = \frac{1}{|G|} \sum_{g' \in G} g'v = \varphi_V(v),$$

since $g' = hg$ runs over G as g does; so $\varphi_V(v) \in V^G$. Thus $\operatorname{Im}\varphi_V = V^G$ and $\varphi_V^2 = \varphi_V$. Then $V = \operatorname{Im}\varphi_V \oplus \ker\varphi_V$: every v is $\varphi_V(v) + (v - \varphi_V(v))$ with $v - \varphi_V(v) \in \ker\varphi_V$, and a vector v in both summands satisfies $v = \varphi_V(v) = 0$. In a basis adapted to this decomposition, φ_V has matrix $\operatorname{diag}(1,\dots,1,0,\dots,0)$ with $\dim V^G$ ones, so $\operatorname{Tr}\varphi_V = \dim V^G$. By linearity of the trace, $\operatorname{Tr}\varphi_V = \frac{1}{|G|}\sum_g \operatorname{Tr}\rho(g) = \frac{1}{|G|}\sum_g \chi_V(g)$. □

Corollary 13.5.2 (Dimension of Hom_G)
For representations V and W of G,

$$\dim \operatorname{Hom}_G(V,W) = \frac{1}{|G|}\sum_{g\in G} \overline{\chi_V(g)}\,\chi_W(g).$$

Proof. By Proposition 13.1.10, $\operatorname{Hom}_G(V,W) = \operatorname{Hom}(V,W)^G$. Apply Proposition 13.5.1 to the representation $\operatorname{Hom}(V,W)$, and use $\chi_{\operatorname{Hom}(V,W)} = \overline{\chi_V}\,\chi_W$ (Proposition 13.4.3). □

This formula suggests an inner product. Let $\mathbb{C}_{\text{class}}(G)$ denote the complex vector space of class functions on G. A class function is determined by its values on the conjugacy classes, and these values can be chosen freely, so the indicator functions of the conjugacy classes form a basis of $\mathbb{C}_{\text{class}}(G)$. In particular, $\dim \mathbb{C}_{\text{class}}(G)$ is the number of conjugacy classes of G.

Definition 13.5.3 (Inner product on class functions)

For class functions α and β on G, let

$$H(\alpha,\beta) = \frac{1}{|G|}\sum_{g\in G} \overline{\alpha(g)}\,\beta(g).$$

This H is a positive-definite Hermitian inner product on $\mathbb{C}_{\text{class}}(G)$. It is conjugate-linear in the first variable and linear in the second, the convention of Chapter 9; it satisfies $H(\beta,\alpha) = \overline{H(\alpha,\beta)}$; and $H(\alpha,\alpha) = \frac{1}{|G|}\sum_g |\alpha(g)|^2 > 0$ for $\alpha \neq 0$. Grouping the terms by conjugacy classes,

$$H(\alpha,\beta) = \frac{1}{|G|}\sum_{C} |C|\,\overline{\alpha(C)}\,\beta(C),$$

where C runs over the conjugacy classes of G and $\alpha(C)$ denotes the value of α on C. This is the form in which we compute with character tables.

Theorem 13.5.4 (Orthogonality of characters)
For representations V and W of G, $H(\chi_V, \chi_W) = \dim \mathrm{Hom}_G(V, W)$. In particular, if V and W are irreducible, then

$$H(\chi_V, \chi_W) = \begin{cases} 1 & \text{if } V \cong W, \\ 0 & \text{if } V \ncong W. \end{cases}$$

So the characters of pairwise non-isomorphic irreducible representations are orthonormal with respect to H.

Proof. The first statement is Corollary 13.5.2, and the second follows from it by Corollary 13.3.2. □

Corollary 13.5.5 (At most as many irreducibles as conjugacy classes)
Up to isomorphism, G has finitely many irreducible representations, and their number is at most the number of conjugacy classes of G.

Proof. Orthonormal vectors are linearly independent: if $\sum_i c_i \chi_{V_i} = 0$, applying $H(\chi_{V_j}, -)$ gives $c_j = 0$. So the characters of pairwise non-isomorphic irreducible representations form a linearly independent subset of $\mathbb{C}_{\mathrm{class}}(G)$, and there are at most $\dim \mathbb{C}_{\mathrm{class}}(G)$ of them. □

From now on, $V_1, \ldots, V_k$ denote the irreducible representations of G, one from each isomorphism class, with characters $\chi_i = \chi_{V_i}$ and dimensions $n_i = \dim V_i$. So far we know that k is at most the number of conjugacy classes; Theorem 13.6.2 below shows that the two numbers are equal.

Corollary 13.5.6 (A representation is determined by its character)
Let W be a representation of G. In any decomposition of W into irreducible subrepresentations, the number of summands isomorphic to V_i is

$$a_i = H(\chi_i, \chi_W) = \dim \mathrm{Hom}_G(V_i, W).$$

So $W \cong \bigoplus_i V_i^{\oplus a_i}$, and the multiplicities a_i do not depend on the decomposition. Two representations with the same character are isomorphic.

Proof. If $W \cong \bigoplus_i V_i^{\oplus a_i}$ (Corollary 13.2.3), then $\chi_W = \sum_i a_i \chi_i$ by Proposition 13.4.3, and $H(\chi_j, \chi_W) = \sum_i a_i H(\chi_j, \chi_i) = a_j$ by Theorem 13.5.4. So the a_i are determined by χ_W, and two representations with the same character have the same multiplicities, hence are isomorphic. □

Corollary 13.5.7 (Irreducibility criterion)
If $W \cong \bigoplus_i V_i^{\oplus a_i}$, then $H(\chi_W, \chi_W) = \sum_i a_i^2$. In particular, W is irreducible if and only if $H(\chi_W, \chi_W) = 1$.

Proof. By orthonormality, $H(\chi_W, \chi_W) = \sum_{i,j} a_i a_j H(\chi_i, \chi_j) = \sum_i a_i^2$. This sum of squares of nonnegative integers equals 1 exactly when one a_i is 1 and the others are 0, that is, when $W \cong V_i$ for some i. □

NOTE The number $H(\chi_W, \chi_W)$ is a quick count of the irreducible pieces of W. For small values it says:

$H(\chi_W, \chi_W)$	W is
1	irreducible;
2	a direct sum of two non-isomorphic irreducibles;
3	a direct sum of three pairwise non-isomorphic irreducibles;
4	a direct sum of four pairwise non-isomorphic irreducibles, or $X \oplus X$ for an irreducible X.

Twisting by a 1-dimensional representation is a cheap way to produce new irreducible representations from old ones.

Proposition 13.5.8 (Twisting by a 1-dimensional representation)
Let L be a 1-dimensional representation of G, given by a homomorphism $\theta : G \to \mathbb{C}^*$. For every representation V of G, $\chi_{V \otimes L} = \theta\, \chi_V$, and $V \otimes L$ is irreducible if and only if V is.

Proof. Fix a basis vector ℓ of L. The map $\iota : V \to V \otimes L$, $v \mapsto v \otimes \ell$, is a linear isomorphism, and

$$g\, \iota(v) = gv \otimes \theta(g)\ell = \iota\big(\theta(g)\, gv\big).$$

So, transported by ι, g acts on $V \otimes L$ as the operator $\theta(g)\rho(g)$ acts on V; taking traces gives $\chi_{V \otimes L} = \theta\, \chi_V$. A subspace of V is invariant under all the operators $\theta(g)\rho(g)$ if and only if it is invariant under all $\rho(g)$, because the $\theta(g)$ are nonzero scalars. Hence ι matches the subrepresentations of V with those of $V \otimes L$, and one is irreducible exactly when the other is. □

13.5.2 The Regular Representation

Definition 13.5.9 (Regular representation)

The ***regular representation*** R of G is the permutation representation for the action of G on itself by left multiplication: $R = \mathbb{C}[G]$ has basis $\{e_h\}_{h \in G}$, and $g\, e_h = e_{gh}$.

Proposition 13.5.10 (Decomposition of the regular representation)

Each irreducible representation V_i occurs in R with multiplicity $n_i = \dim V_i$: $R \cong \bigoplus_i V_i^{\oplus n_i}$.

Proof. By Proposition 13.4.3(v), $\chi_R(g)$ is the number of $h \in G$ with $gh = h$, which is $|G|$ if $g = e$ and 0 otherwise. By Corollary 13.5.6, the multiplicity of V_i in R is

$$H(\chi_i, \chi_R) = \frac{1}{|G|} \overline{\chi_i(e)}\, |G| = \dim V_i.$$

□

Corollary 13.5.11 (Sum of squares of the dimensions)

$\sum_{i=1}^{k} n_i^2 = |G|$.

Proof. Compare dimensions in Proposition 13.5.10: $|G| = \dim R = \sum_i n_i \dim V_i = \sum_i n_i^2$. □

13.6 Characters Form an Orthonormal Basis

We know that $\chi_1, \dots, \chi_k$ are orthonormal in $\mathbb{C}_{\text{class}}(G)$, so they are linearly independent and k is at most the number of conjugacy classes. We now prove that they also span $\mathbb{C}_{\text{class}}(G)$, so that k equals the number of conjugacy classes. The proof uses a weighted version of the averaging projection.

13.6.1 Weighted Averages

Proposition 13.6.1 (Weighted averaging operators)

Let α be a class function on G and V a representation of G, and let

$$\varphi_{\alpha, V} = \frac{1}{|G|} \sum_{g \in G} \alpha(g)\, g : V \to V.$$

Then $\varphi_{\alpha,V}$ is a homomorphism of representations, and $\operatorname{Tr}\varphi_{\alpha,V} = \frac{1}{|G|}\sum_{g\in G}\alpha(g)\,\chi_V(g)$.

Proof. For $h \in G$,

$$h\,\varphi_{\alpha,V}\,h^{-1} = \frac{1}{|G|}\sum_{g\in G}\alpha(g)\,hgh^{-1} = \frac{1}{|G|}\sum_{g'\in G}\alpha(h^{-1}g'h)\,g'$$
$$= \frac{1}{|G|}\sum_{g'\in G}\alpha(g')\,g' = \varphi_{\alpha,V},$$

where we substituted $g' = hgh^{-1}$, which runs over G as g does, and used that α is a class function. So $h\,\varphi_{\alpha,V} = \varphi_{\alpha,V}\,h$ for every $h \in G$. The trace formula follows from linearity of the trace. □

For the constant function $\alpha = 1$, $\varphi_{1,V}$ is the averaging projection φ_V of Proposition 13.5.1. We also note two simple facts. If $V = X \oplus Y$ is a direct sum of subrepresentations, then $\varphi_{\alpha,V}$ preserves X and Y and restricts to $\varphi_{\alpha,X}$ and $\varphi_{\alpha,Y}$ on them, because each g does. If $\psi : X \to Y$ is an isomorphism of representations, then $\varphi_{\alpha,Y} = \psi\,\varphi_{\alpha,X}\,\psi^{-1}$, because $\rho_Y(g) = \psi\,\rho_X(g)\,\psi^{-1}$ for each g.

Theorem 13.6.2 (Irreducible characters form an orthonormal basis)

The characters $\chi_1, \ldots, \chi_k$ of the irreducible representations of G form an orthonormal basis of $\mathbb{C}_{\text{class}}(G)$ with respect to H. Every class function β on G can be written as

$$\beta = \sum_{i=1}^{k} H(\chi_i, \beta)\,\chi_i.$$

In particular, the number of irreducible representations of G, up to isomorphism, equals the number of conjugacy classes of G.

Proof. Orthonormality is Theorem 13.5.4. We first show that a class function α with $H(\alpha, \chi_i) = 0$ for all i is zero.

The complex conjugate $\overline{\alpha}$ is also a class function. By Proposition 13.6.1, $\varphi_{\overline{\alpha},V_i}$ is a homomorphism of representations $V_i \to V_i$, so by Schur's Lemma it equals $c\,\mathrm{id}_{V_i}$ for some $c \in \mathbb{C}$. Taking traces,

$$n_i c = \operatorname{Tr}\varphi_{\overline{\alpha},V_i} = \frac{1}{|G|}\sum_{g\in G}\overline{\alpha(g)}\,\chi_i(g) = H(\alpha, \chi_i) = 0,$$

so $\varphi_{\overline{\alpha},V_i} = 0$ for every i. Every representation V is a direct sum of subrepresentations, each isomorphic to some V_i (Corollary 13.2.3), so by the two facts above,

$\varphi_{\overline{\alpha},V} = 0$ for every representation V. Take V to be the regular representation R, and apply $\varphi_{\overline{\alpha},R}$ to the basis vector e_e indexed by the identity element:

$$0 = \varphi_{\overline{\alpha},R}(e_e) = \frac{1}{|G|} \sum_{g \in G} \overline{\alpha(g)}\, e_g.$$

Since the e_g are linearly independent, $\alpha(g) = 0$ for every $g \in G$.

Now let β be any class function, and let $\alpha = \beta - \sum_i H(\chi_i, \beta)\, \chi_i$. For each j, orthonormality gives $H(\chi_j, \alpha) = H(\chi_j, \beta) - H(\chi_j, \beta) = 0$, hence $H(\alpha, \chi_j) = \overline{H(\chi_j, \alpha)} = 0$. By what we have shown, $\alpha = 0$, which is the stated formula for β. So $\chi_1, \ldots, \chi_k$ span $\mathbb{C}_{\text{class}}(G)$ and form an orthonormal basis of it, and $k = \dim \mathbb{C}_{\text{class}}(G)$ is the number of conjugacy classes. □

Corollary 13.6.3 (Square character tables)

The character table of a finite group is a square matrix.

NOTE By Theorem 3.4.6, two permutations in S_n are conjugate exactly when they have the same cycle type, and cycle types correspond to partitions of n. So the number of irreducible representations of S_n is $p(n)$, the number of partitions of n. For instance $p(3) = 3$, in agreement with Proposition 13.3.5, while $p(4) = 5$ and $p(5) = 7$.

13.6.2 Projection onto Isotypic Components

Let W be a representation with a decomposition $W = \bigoplus_j X_j$ into irreducible subrepresentations. The ***isotypic component*** of W of type V_i is the sum of those X_j that are isomorphic to V_i; it is isomorphic to $V_i^{\oplus a_i}$, where a_i is the multiplicity of V_i in W. Weighted averages give an explicit projection onto it.

For irreducible V_i and V_j, take $\alpha = \overline{\chi_i}$ in Proposition 13.6.1. As in the proof of Theorem 13.6.2, $\varphi_{\overline{\chi_i},V_j}$ is a scalar c with $n_j c = H(\chi_i, \chi_j)$, so

$$\varphi_{\overline{\chi_i},V_j} = \frac{1}{n_j} H(\chi_i, \chi_j)\, \mathrm{id}_{V_j} = \begin{cases} \dfrac{1}{n_i}\, \mathrm{id}_{V_i} & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Proposition 13.6.4 (Projection onto an isotypic component)

Let W be a representation of G, and let

$$\pi_i = \frac{n_i}{|G|} \sum_{g \in G} \overline{\chi_i(g)}\, g : W \to W.$$

Then π_i is the identity on every subrepresentation of W isomorphic to V_i, and zero on every irreducible subrepresentation not isomorphic to V_i. Hence, for any decomposition of W into irreducible subrepresentations, π_i is the projection onto the isotypic component of type V_i along the sum of the other isotypic components. In particular, this isotypic component is $\operatorname{Im}\pi_i$, so it does not depend on the decomposition.

Proof. We have $\pi_i = n_i\,\varphi_{\overline{\chi_i},W}$. On an irreducible subrepresentation $X \cong V_j$, π_i acts as $n_i\,\varphi_{\overline{\chi_i},X}$, which is conjugate to $n_i\,\varphi_{\overline{\chi_i},V_j}$. By the computation above, this is the identity if $j = i$ and zero otherwise. □

NOTE When V_i is the trivial representation, $n_i = 1$ and $\chi_i = 1$, and π_i is the averaging projection φ_W onto W^G of Proposition 13.5.1.

13.6.3 Column Orthogonality

Orthonormality of the irreducible characters says that the rows of the character table are orthonormal, once the columns are weighted by the sizes of the classes. Because the table is square, its columns are orthogonal as well.

Proposition 13.6.5 (Column orthogonality)
Let $C_1, \dots, C_k$ be the conjugacy classes of G, with representatives $g_1, \dots, g_k$. Then

$$\sum_{i=1}^{k} \chi_i(g_s)\,\overline{\chi_i(g_t)} = \begin{cases} \dfrac{|G|}{|C_s|} & \text{if } s = t, \\ 0 & \text{if } s \neq t. \end{cases}$$

Proof. Let X be the $k \times k$ matrix with entries $X_{is} = \chi_i(g_s)$, which is square by Theorem 13.6.2, and let $D = \operatorname{diag}(|C_1|/|G|, \dots, |C_k|/|G|)$. Orthonormality of the characters, written class by class, says

$$\sum_{s=1}^{k} \overline{X_{is}}\,\frac{|C_s|}{|G|}\,X_{js} = \delta_{ij}, \qquad \text{that is,} \qquad \overline{X}\,D\,X^{t} = I.$$

So the square matrix $\overline{X}D$ has inverse X^t, and therefore also $X^t\,\overline{X}D = I$, that is, $X^t\,\overline{X} = D^{-1}$. The (s,t) entry of this identity is $\sum_i X_{is}\,\overline{X_{it}} = \delta_{st}\,|G|/|C_s|$. □

NOTE By Proposition 3.2.13, $|C_s| = [G : Z(g_s)]$, so $|G|/|C_s|$ is the order of the centralizer $Z(g_s) = \{h \in G : hg_s = g_sh\}$. Column orthogonality gives an independent check on a computed character table. It also shows that

characters separate conjugacy classes: if $\chi_i(g) = \chi_i(h)$ for every i, then $\sum_i \chi_i(g)\overline{\chi_i(h)} = \sum_i |\chi_i(g)|^2 > 0$, so g and h lie in the same conjugacy class.

Exercise 13.6.6 (Representations and Characters)

(1) Decompose the permutation representation $\mathbb{C}^3$ of S_3 into irreducibles using the inner product H.
(2) Let V be the standard representation of S_3. Use characters and the inner product H to decompose $V \otimes V$, $\operatorname{Sym}^2 V$ and $\bigwedge^2 V$ into irreducibles, and compare with Example 13.3.6.
(3) Compute the character table of the Klein four-group $\mathbb{Z}/2 \times \mathbb{Z}/2$. How many irreducible representations does it have, and what are their dimensions?
(4) Let G act on a finite set S, and let $\mathbb{C}[S]$ be the permutation representation.
   (a) Show that $\dim \mathbb{C}[S]^G$ is the number of orbits of G on S, and deduce Burnside's lemma (Theorem 11.1.3) from Proposition 13.5.1.
   (b) Suppose that $|S| \geqslant 2$ and that G acts *doubly transitively*: for any two pairs (s,t) and (s',t') of distinct elements of S, there is $g \in G$ with $gs = s'$ and $gt = t'$. Show that $\mathbb{C}[S] \cong \mathbb{C} \oplus V_S$, where $\mathbb{C}$ is the trivial representation and $V_S = \{\sum_s c_s e_s : \sum_s c_s = 0\}$ is irreducible. Conclude that the standard representation of S_n is irreducible for every $n \geqslant 2$.

Solution. **(1)** The permutation character counts fixed points: $\chi(e) = 3$, $\chi((12)) = 1$, $\chi((123)) = 0$. With the class sizes $1, 3, 2$,

$$H(\chi_U, \chi) = \tfrac{1}{6}(3+3+0) = 1, \qquad H(\chi_{U'}, \chi) = \tfrac{1}{6}(3-3+0) = 0,$$

and $H(\chi_V, \chi) = \frac{1}{6}(6+0+0) = 1$. So $\mathbb{C}^3 \cong U \oplus V$.
(2) The character of $V \otimes V$ is $\chi_V^2 = (4, 0, 1)$, and

$$H(\chi_U, \chi_V^2) = \tfrac{1}{6}(4+0+2) = 1, \qquad H(\chi_{U'}, \chi_V^2) = \tfrac{1}{6}(4+0+2) = 1,$$

and $H(\chi_V, \chi_V^2) = \frac{1}{6}(8+0-2) = 1$, so $V \otimes V \cong U \oplus U' \oplus V$. For the symmetric and exterior squares we need $\chi_V(g^2)$: the squares of e, (12), (123) are e, e, (132), so $\chi_V(g^2) = (2, 2, -1)$. By Proposition 13.4.3(vi),

$$\chi_{\operatorname{Sym}^2 V} = \tfrac{1}{2}(4+2,\ 0+2,\ 1-1) = (3, 1, 0),$$
$$\chi_{\bigwedge^2 V} = \tfrac{1}{2}(4-2,\ 0-2,\ 1+1) = (1, -1, 1).$$

The first is $\chi_U + \chi_V$ and the second is $\chi_{U'}$, so $\mathrm{Sym}^2 V \cong U \oplus V$ and $\bigwedge^2 V \cong U'$, as in Example 13.3.6.

(3) The group is abelian of order 4, so each element is its own conjugacy class, and there are four irreducible representations, all 1-dimensional (Proposition 13.1.7); indeed $4 \cdot 1^2 = 4$. They are the elements of the dual group: for $(a, b) \in \{0,1\}^2$, the character $\chi_{ab}(x, y) = (-1)^{ax+by}$.

	$(0,0)$	$(1,0)$	$(0,1)$	$(1,1)$
χ_{00}	1	1	1	1
χ_{10}	1	-1	1	-1
χ_{01}	1	1	-1	-1
χ_{11}	1	-1	-1	1

(4) (a) A vector $\sum_s c_s e_s$ is invariant if and only if $c_{gs} = c_s$ for all g and s, that is, if and only if $s \mapsto c_s$ is constant on each orbit. So the sums $\sum_{s \in O} e_s$, one for each orbit O, form a basis of $\mathbb{C}[S]^G$. By Proposition 13.5.1 and Proposition 13.4.3(v), the number of orbits is therefore $\frac{1}{|G|}\sum_{g \in G} |S^g|$, which is Burnside's lemma.

(b) The line spanned by $\sum_s e_s$ is a copy of the trivial representation, V_S is a subrepresentation complementary to it, and $\chi_{V_S}(g) = |S^g| - 1$. The action is transitive (given s, s', choose $t \neq s$ and $t' \neq s'$, and use double transitivity), so by (a), $\frac{1}{|G|}\sum_g |S^g| = 1$. The diagonal $\{(s,s)\}$ and its complement are the orbits of G on $S \times S$, and $|(S \times S)^g| = |S^g|^2$; so by (a) applied to $S \times S$, $\frac{1}{|G|}\sum_g |S^g|^2 = 2$. Since χ_{V_S} is real,

$$H(\chi_{V_S}, \chi_{V_S}) = \frac{1}{|G|}\sum_{g \in G} \big(|S^g| - 1\big)^2 = 2 - 2 \cdot 1 + 1 = 1,$$

and V_S is irreducible by Corollary 13.5.7. The group S_n acts doubly transitively on $\{1, \ldots, n\}$ for $n \geqslant 2$, and V_S is then the standard representation.

CHAPTER 14

Character Tables and the Representation Ring

With the results of Chapter 13 we can compute the character table of a given group without first constructing all of its representations. We use the notation of that chapter: $V_1, \dots, V_k$ are the irreducible representations of G, χ_i their characters, and H the inner product on class functions (Definition 13.5.3). The tools are these.

- The number of irreducible representations equals the number of conjugacy classes (Theorem 13.6.2), and the squares of their dimensions add up to $|G|$ (Corollary 13.5.11).
- The irreducible characters are orthonormal (Theorem 13.5.4), and the columns of the table are orthogonal (Proposition 13.6.5).
- A representation W is irreducible exactly when $H(\chi_W, \chi_W) = 1$ (Corollary 13.5.7), and $H(\chi_i, \chi_W)$ is the multiplicity of V_i in W (Corollary 13.5.6).
- New representations come from old ones: permutation representations, twists by 1-dimensional representations (Proposition 13.5.8), tensor products, symmetric and exterior squares, whose characters are given by Proposition 13.4.3.

We work out the character tables of S_4, A_4, S_5 and A_5 in this way; the table of S_3 is Example 13.4.5. The chapter ends with the representation ring, which packages the operations $\oplus$ and $\otimes$ into a ring.

For a symmetric group, U, U' and V denote its trivial, alternating and standard representations (Example 13.1.4), and $\chi_V(\sigma) = |\{i : \sigma(i) = i\}| - 1$, because $U \oplus V$ is the permutation representation, whose character counts fixed points. In the tables, each column is labelled by a representative of a conjugacy class, with the size of the class below it. When we write a class function as a vector, its entries are its values on the classes in the order of the table.

Two general facts about 1-dimensional representations will be used repeatedly. First, if L is 1-dimensional and X is irreducible, then $X \otimes L$ is irreducible with character $\chi_X\chi_L$ (Proposition 13.5.8). Second, S_n ($n \geqslant 2$) has exactly two 1-dimensional representations, U and U'. Indeed, a homomorphism $\theta : S_n \to \mathbb{C}^*$ takes the same

value c on all transpositions, since they are conjugate (Theorem 3.4.6) and $\mathbb{C}^*$ is abelian; and $c^2 = \theta(e) = 1$. As the transpositions generate S_n (Proposition 3.4.9), θ is the constant function 1 if $c = 1$, and $\theta = \operatorname{sgn}$ if $c = -1$.

14.1 The Character Table of S_4

By Theorem 3.4.6, the conjugacy classes of S_4 are given by cycle type: the identity, the six transpositions, the three double transpositions such as $(12)(34)$, the eight 3-cycles and the six 4-cycles. So S_4 has five irreducible representations, and the squares of their dimensions add up to 24.

We know three of them. The rows of U and U' are the constant function 1 and sgn, and $\chi_V = (3, 1, -1, 0, -1)$ in the order of the table below. Since

$$H(\chi_V, \chi_V) = \tfrac{1}{24}\big(9 + 6\cdot 1 + 3\cdot 1 + 8\cdot 0 + 6\cdot 1\big) = 1,$$

V is irreducible. The twist $V' = V \otimes U'$ is irreducible with character $\chi_V \operatorname{sgn} = (3, -1, -1, 0, 1)$, and $V' \ncong V$ because the characters differ at (12). So far $1 + 1 + 9 + 9 = 20$, so the fifth irreducible representation W has dimension 2.

To find χ_W, observe that $W \otimes U'$ is also a 2-dimensional irreducible representation, and W is the only one; so $W \otimes U' \cong W$ and $\chi_W \operatorname{sgn} = \chi_W$. Hence χ_W vanishes on the odd classes, those of (12) and (1234). Write $b = \chi_W((12)(34))$ and $a = \chi_W((123))$. Orthogonality to χ_U and to χ_V gives

$$H(\chi_U, \chi_W) = \tfrac{1}{24}(2 + 3b + 8a) = 0, \qquad H(\chi_V, \chi_W) = \tfrac{1}{24}(6 - 3b) = 0,$$

so $b = 2$ and $a = -1$. The table is complete:

	e	(12)	$(12)(34)$	(123)	(1234)
size	1	6	3	8	6
U	1	1	1	1	1
U'	1	-1	1	1	-1
W	2	0	2	-1	0
V	3	1	-1	0	-1
V'	3	-1	-1	0	1

As a check, $H(\chi_W, \chi_W) = \frac{1}{24}(4 + 3\cdot 4 + 8\cdot 1) = 1$.

14.1.1 Identifying W

The value $\chi_W((12)(34)) = 2$ tells us a great deal. The operator by which $(12)(34)$ acts on W has order dividing 2, so it is diagonalizable with eigenvalues ± 1 (Lemma 13.1.5), and two such eigenvalues add up to 2 only if both are 1. So

$(12)(34)$ acts on W as the identity, and so do the other two double transpositions, which have the same character value. Hence the subgroup

$$K = \{e,\ (12)(34),\ (13)(24),\ (14)(23)\} \cong \mathbb{Z}/2 \times \mathbb{Z}/2$$

lies in the kernel of $S_4 \to \mathrm{GL}(W)$. It is normal, being the union of the conjugacy classes of e and of $(12)(34)$ (Proposition 3.2.13), so this homomorphism factors through S_4/K (Theorem 3.3.2).

To identify S_4/K, let S_4 act on the three ways of splitting $\{1,2,3,4\}$ into two pairs,

$$P_1 = \{12 \mid 34\}, \qquad P_2 = \{13 \mid 24\}, \qquad P_3 = \{14 \mid 23\}.$$

This gives a homomorphism $q : S_4 \to S_3$. Each double transposition preserves all three splittings, so $K \subseteq \ker q$. The transposition (12) fixes P_1 and swaps P_2 and P_3, and (123) permutes $P_1 \to P_3 \to P_2 \to P_1$, so the image of q contains elements of orders 2 and 3, and q is surjective. Then $|\ker q| = 24/6 = 4$, so $\ker q = K$, and $S_4/K \cong S_3$ by the First Isomorphism Theorem (Theorem 3.3.3). Under q, transpositions and 4-cycles go to transpositions (for instance, (1234) fixes P_2 and swaps P_1 and P_3), and 3-cycles go to 3-cycles.

Now let W_0 be the standard representation of S_3, regarded as a representation of S_4 in which g acts as $q(g)$ does. Its character is $\chi_{W_0}(g) = \chi_{V_{S_3}}(q(g))$, where V_{S_3} is the standard representation of S_3; by Example 13.4.5 this is $(2,0,2,-1,0)$, which is χ_W. So $W \cong W_0$ (Corollary 13.5.6): W is the standard representation of S_3, pulled back to S_4 along $S_4 \to S_4/K \cong S_3$.

14.1.2 A Second Construction of W

We can also find W inside $V \otimes V$. Its character is $\chi_V^2 = (9,1,1,0,1)$, and

$$\begin{aligned}
H(\chi_U, \chi_V^2) &= \tfrac{1}{24}(9+6+3+0+6) = 1,\\
H(\chi_{U'}, \chi_V^2) &= \tfrac{1}{24}(9-6+3+0-6) = 0,\\
H(\chi_V, \chi_V^2) &= \tfrac{1}{24}(27+6-3+0-6) = 1,\\
H(\chi_{V'}, \chi_V^2) &= \tfrac{1}{24}(27-6-3+0+6) = 1.
\end{aligned}$$

So $V \otimes V$ contains $U \oplus V \oplus V'$, of dimension 7, and the remaining summand has dimension 2 and contains none of U, U', V, V'; it must be W. Thus

$$V \otimes V \cong U \oplus V \oplus V' \oplus W,$$

and $\chi_W = \chi_V^2 - \chi_U - \chi_V - \chi_{V'}$, which again gives $(2,0,2,-1,0)$. (Consistently, $H(\chi_V^2, \chi_V^2) = \frac{1}{24}(81+6+3+0+6) = 4$, the value for four non-isomorphic summands.)

14.2 The Character Table of A_4

The alternating group A_4 consists of the identity, the three double transpositions and the eight 3-cycles. Its conjugacy classes are found with the splitting criterion for conjugacy classes in A_n (Proposition 11.4.3): a conjugacy class of S_n contained in A_n remains a single conjugacy class of A_n if some odd permutation commutes with one of its elements, and otherwise it splits into two classes of equal size. The transposition (12) commutes with $(12)(34)$, so the double transpositions form one class of A_4. The centralizer of (123) in S_4 has order $24/8 = 3$ (Proposition 3.2.13), so it is $\langle(123)\rangle \subseteq A_4$, and the eight 3-cycles split into the two classes

$$\{(123),\ (134),\ (243),\ (142)\} \qquad \text{and} \qquad \{(132),\ (143),\ (234),\ (124)\}.$$

So A_4 has four conjugacy classes, and therefore four irreducible representations, the squares of whose dimensions add up to 12.

The 1-dimensional representations. The subgroup K of Section 14.1, consisting of e and the three double transpositions, is a union of conjugacy classes, hence normal in A_4 (Proposition 3.2.13), and A_4/K has order 3, so it is cyclic, generated by the coset of (123). Every homomorphism $\theta : A_4 \to \mathbb{C}^*$ is trivial on K: the three double transpositions are conjugate, so θ takes a common value ε on them, with $\varepsilon^2 = 1$; and $(12)(34)\,(13)(24) = (14)(23)$ gives $\varepsilon^2 = \varepsilon$, so $\varepsilon = 1$. So θ factors through $A_4/K \cong \mathbb{Z}/3$, and with $\omega = e^{2\pi i/3}$ we obtain exactly three 1-dimensional representations U, U', U'', sending (123) to 1, ω, ω^2 respectively. (Here U' and U'' are new; the alternating representation of S_4 becomes trivial on A_4.) Since $(132) = (123)^2$, their values on (132) are 1, ω^2, ω.

The last representation. The fourth irreducible representation has dimension $\sqrt{12-3} = 3$. The standard representation V of S_4, with A_4 acting through the inclusion $A_4 \subset S_4$, has character $(3, -1, 0, 0)$ on A_4, and

$$H(\chi_V, \chi_V) = \tfrac{1}{12}\big(9 + 3\cdot 1\big) = 1,$$

so V remains irreducible on A_4. It is the fourth irreducible representation.

	e	$(12)(34)$	(123)	(132)
size	1	3	4	4
U	1	1	1	1
U'	1	1	ω	ω^2
U''	1	1	ω^2	ω
V	3	-1	0	0

NOTE The representation W of S_4 does not stay irreducible on A_4. Its character on A_4 is $(2, 2, -1, -1)$, which equals $\chi_{U'} + \chi_{U''}$ because $\omega + \omega^2 = -1$;

so on A_4, W is isomorphic to $U' \oplus U''$. This also follows from the description of W above: K acts trivially, and A_4/K maps onto the subgroup $A_3 \cong \mathbb{Z}/3$ of S_3, whose generator acts on the standard representation of S_3 with eigenvalues ω and ω^2 (Lemma 13.3.4).

14.3 The Character Table of S_5

The group S_5 has seven conjugacy classes, one for each cycle type:

cycle type	1^5	$2 \cdot 1^3$	$2^2 \cdot 1$	$3 \cdot 1^2$	$3 \cdot 2$	$4 \cdot 1$	5
representative	e	(12)	$(12)(34)$	(123)	$(123)(45)$	(1234)	(12345)
size	1	10	15	20	20	30	24

So it has seven irreducible representations, and the squares of their dimensions add up to 120. We start from the ones we know.

- U and U', the only 1-dimensional representations.
- The standard representation V, of dimension 4, with $\chi_V = (4, 2, 0, 1, -1, 0, -1)$. It is irreducible, since
$$H(\chi_V, \chi_V) = \tfrac{1}{120}\big(16 + 10 \cdot 4 + 0 + 20 \cdot 1 + 20 \cdot 1 + 0 + 24 \cdot 1\big) = 1.$$
- The twist $V' = V \otimes U'$, irreducible of dimension 4, with $\chi_{V'} = \chi_V \operatorname{sgn} = (4, -2, 0, 1, 1, 0, -1)$. It is not isomorphic to V, since the characters differ at (12).

These account for $1 + 1 + 16 + 16 = 34$, so the remaining three irreducible representations have dimensions $d_1 \leqslant d_2 \leqslant d_3$ with $d_1^2 + d_2^2 + d_3^2 = 86$, and $d_1 \geqslant 2$ because U and U' are the only 1-dimensional ones. Then $3d_3^2 \geqslant 86$ and $d_3^2 \leqslant 86 - 8$, so $d_3 \in \{6, 7, 8\}$; this leaves 22, 37 or 50 for $d_1^2 + d_2^2$, and the only way to write one of these as a sum of two squares of integers $\geqslant 2$ is $50 = 5^2 + 5^2$. So the remaining dimensions are 5, 5 and 6.

14.3.1 Finding the Remaining Irreducibles

The most effective way to find them is to decompose $V \otimes V$, of dimension 16, or rather its two pieces $\operatorname{Sym}^2 V$ and $\bigwedge^2 V$, of dimensions 10 and 6. By Proposition 13.4.3(vi),

$$\chi_{\operatorname{Sym}^2 V}(g) = \tfrac{1}{2}\big(\chi_V(g)^2 + \chi_V(g^2)\big), \qquad \chi_{\bigwedge^2 V}(g) = \tfrac{1}{2}\big(\chi_V(g)^2 - \chi_V(g^2)\big),$$

which come from the fact that if g has eigenvalues $\lambda_1, \dots, \lambda_4$ on V, then it has eigenvalues $\lambda_i\lambda_j$, $i \leqslant j$, on $\operatorname{Sym}^2 V$ and $\lambda_i\lambda_j$, $i < j$, on $\bigwedge^2 V$. To use them we need $\chi_V(g^2)$. Squaring the representatives: $(12)^2 = ((12)(34))^2 = e$; $(123)^2 = (132)$

and $((123)(45))^2 = (132)$ are 3-cycles; $(1234)^2 = (13)(24)$ is a double transposition; and $(12345)^2 = (13524)$ is a 5-cycle. This gives:

	e	(12)	$(12)(34)$	(123)	$(123)(45)$	(1234)	(12345)
size	1	10	15	20	20	30	24
$\chi_V(g)$	4	2	0	1	-1	0	-1
$\chi_V(g^2)$	4	4	4	1	1	0	-1
$\chi_{\bigwedge^2 V}$	6	0	-2	0	0	0	1
$\chi_{\mathrm{Sym}^2 V}$	10	4	2	1	1	0	0

First, $\bigwedge^2 V$ is irreducible:

$$H(\chi_{\bigwedge^2 V}, \chi_{\bigwedge^2 V}) = \tfrac{1}{120}\big(36 + 0 + 15\cdot 4 + 0 + 0 + 0 + 24\cdot 1\big) = \tfrac{120}{120} = 1.$$

It is the 6-dimensional irreducible representation. Next,

$$H(\chi_{\mathrm{Sym}^2 V}, \chi_{\mathrm{Sym}^2 V}) = \tfrac{1}{120}\big(100 + 10\cdot 16 + 15\cdot 4 + 20\cdot 1 + 20\cdot 1 + 0 + 0\big) = \tfrac{360}{120} = 3,$$

so $\mathrm{Sym}^2 V$ is a direct sum of three pairwise non-isomorphic irreducible representations. Two of them are U and V:

$$H(\chi_U, \chi_{\mathrm{Sym}^2 V}) = \tfrac{1}{120}(10 + 40 + 30 + 20 + 20 + 0 + 0) = 1,$$
$$H(\chi_V, \chi_{\mathrm{Sym}^2 V}) = \tfrac{1}{120}(40 + 80 + 0 + 20 - 20 + 0 + 0) = 1.$$

So $\mathrm{Sym}^2 V \cong U \oplus V \oplus W$, where W is irreducible of dimension $10 - 1 - 4 = 5$, with character

$$\chi_W = \chi_{\mathrm{Sym}^2 V} - \chi_U - \chi_V = (5, 1, 1, -1, 1, -1, 0).$$

Finally, $W' = W \otimes U'$ is irreducible with character $\chi_W \operatorname{sgn} = (5, -1, 1, -1, -1, 1, 0)$, and $W' \ncong W$ because the characters differ at (12). This gives all seven irreducible representations.

14.3.2 The Complete Character Table

	e	(12)	$(12)(34)$	(123)	$(123)(45)$	(1234)	(12345)
size	1	10	15	20	20	30	24
U	1	1	1	1	1	1	1
U'	1	-1	1	1	-1	-1	1
V	4	2	0	1	-1	0	-1
V'	4	-2	0	1	1	0	-1
W	5	1	1	-1	1	-1	0
W'	5	-1	1	-1	-1	1	0
$\bigwedge^2 V$	6	0	-2	0	0	0	1

As checks, $1+1+16+16+25+25+36 = 120$, and column orthogonality holds; for instance, the column of (12) gives $1+1+4+4+1+1+0 = 12 = 120/10$.

NOTE The exterior powers $\bigwedge^k V$ of the standard representation of S_5 are $\bigwedge^0 V = U$, $\bigwedge^1 V = V$, the irreducible $\bigwedge^2 V$, $\bigwedge^3 V \cong V'$ and $\bigwedge^4 V \cong U'$ (see the exercises), so all of them are irreducible. This is a general fact: for the standard representation V of S_n, every $\bigwedge^k V$ with $0 \leqslant k \leqslant n-1$ is irreducible (Fulton and Harris, *Representation Theory: A First Course*, §3.2).

14.4 The Character Table of A_5

By the splitting criterion for conjugacy classes in A_n (Proposition 11.4.3), the conjugacy classes of A_5 are as follows. The transposition (12) commutes with $(12)(34)$, and (45) commutes with (123), so the double transpositions and the 3-cycles each remain a single class of A_5. The centralizer of (12345) in S_5 has order $120/24 = 5$ (Proposition 3.2.13), so it is $\langle(12345)\rangle \subseteq A_5$, and the twenty-four 5-cycles split into two classes of twelve. Conjugating (12345) by the odd permutation (45) gives (12354), which lies in the other class: if $(45)(12345)(45)^{-1} = a(12345)a^{-1}$ with $a \in A_5$, then the odd permutation $a^{-1}(45)$ would commute with (12345). So A_5 has five conjugacy classes:

representative	e	$(12)(34)$	(123)	(12345)	(12354)
size	1	15	20	12	12

It therefore has five irreducible representations.

Letting only the even permutations act turns every representation of S_5 into a representation of A_5, its *restriction* (restriction is studied systematically in Chapter 15). Elements of A_5 act trivially on U', so U' restricts to the trivial representation, and for any representation X of S_5, the restrictions of X and $X \otimes U'$ to A_5 have the same character, hence are isomorphic. So it suffices to restrict U, V, W and $\bigwedge^2 V$:

	e	$(12)(34)$	(123)	(12345)	(12354)
size	1	15	20	12	12
U	1	1	1	1	1
V	4	0	1	-1	-1
W	5	1	-1	0	0
$\bigwedge^2 V$	6	-2	0	1	1

Computing inner products over A_5, $H(\chi_V, \chi_V) = \frac{1}{60}(16+0+20+12+12) = 1$ and $H(\chi_W, \chi_W) = \frac{1}{60}(25+15+20+0+0) = 1$, so V and W remain irreducible.

However,

$$H(\chi_{\bigwedge^2 V}, \chi_{\bigwedge^2 V}) = \tfrac{1}{60}(36 + 15 \cdot 4 + 0 + 12 + 12) = 2,$$

so on A_5, $\bigwedge^2 V$ is a direct sum of two non-isomorphic irreducible representations. Neither is U, V or W, since

$$\begin{aligned} H(\chi_U, \chi_{\bigwedge^2 V}) &= \tfrac{1}{60}(6 - 30 + 0 + 12 + 12) = 0, \\ H(\chi_V, \chi_{\bigwedge^2 V}) &= \tfrac{1}{60}(24 + 0 + 0 - 12 - 12) = 0, \\ H(\chi_W, \chi_{\bigwedge^2 V}) &= \tfrac{1}{60}(30 - 30 + 0 + 0 + 0) = 0. \end{aligned}$$

So $\bigwedge^2 V \cong Y \oplus Z$ on A_5, where Y and Z are the two remaining irreducible representations of A_5. Their dimensions satisfy $1+16+25+(\dim Y)^2+(\dim Z)^2 = 60$ and $\dim Y + \dim Z = 6$, so $\dim Y = \dim Z = 3$.

14.4.1 Determining Y and Z

We know $\chi_Y + \chi_Z = \chi_{\bigwedge^2 V} = (6, -2, 0, 1, 1)$. Write $\chi_Y - \chi_Z = (0, f, b, c, d)$; the first entry is $\dim Y - \dim Z = 0$. Since Y and Z are not isomorphic to U, V or W, the class function $\chi_Y - \chi_Z$ is orthogonal to χ_U, χ_V and χ_W, which are real-valued:

$$15f + 20b + 12c + 12d = 0, \qquad 20b - 12c - 12d = 0, \qquad 15f - 20b = 0.$$

Adding the first two equations gives $15f + 40b = 0$, and with the third, $b = f = 0$ and then $d = -c$. So $\chi_Y - \chi_Z = (0, 0, 0, a, -a)$ for some $a \in \mathbb{C}$. Since $H(\chi_Y - \chi_Z, \chi_Y - \chi_Z) = H(\chi_Y, \chi_Y) + H(\chi_Z, \chi_Z) = 2$,

$$\tfrac{1}{60}\big(12|a|^2 + 12|a|^2\big) = 2, \qquad \text{so} \qquad |a|^2 = 5.$$

This does not yet determine a. But every character of A_5 is real-valued. Indeed, every element of A_5 is conjugate in A_5 to its inverse: a 3-cycle or a double transposition has the same cycle type as its inverse, and these classes of S_5 do not split in A_5; and a 5-cycle $(a_1\ a_2\ a_3\ a_4\ a_5)$ has inverse $(a_1\ a_5\ a_4\ a_3\ a_2)$, which is its conjugate by the even permutation $(a_2\ a_5)(a_3\ a_4)$. So for every representation X of A_5 and every $g \in A_5$, $\chi_X(g) = \chi_X(g^{-1}) = \overline{\chi_X(g)}$ by Proposition 13.4.3(iii). Hence a is real and $a = \pm\sqrt{5}$. The sign only decides which of the two representations we call Y; taking $a = \sqrt{5}$, we get $\chi_Y = \frac{1}{2}\big((\chi_Y + \chi_Z) + (\chi_Y - \chi_Z)\big)$, and similarly for χ_Z:

	e	$(12)(34)$	(123)	(12345)	(12354)
size	1	15	20	12	12
U	1	1	1	1	1
Y	3	-1	0	$\frac{1+\sqrt{5}}{2}$	$\frac{1-\sqrt{5}}{2}$
Z	3	-1	0	$\frac{1-\sqrt{5}}{2}$	$\frac{1+\sqrt{5}}{2}$
V	4	0	1	-1	-1
W	5	1	-1	0	0

The golden ratio $\frac{1+\sqrt{5}}{2}$ appears naturally. The representations Y and Z are exchanged by the automorphism $g \mapsto \tau g \tau^{-1}$ of A_5, for a transposition $\tau \in S_5$: letting $g \in A_5$ act on Y as $\tau g \tau^{-1}$ does defines a new representation with character $g \mapsto \chi_Y(\tau g \tau^{-1})$. Conjugation by τ fixes the classes of e, $(12)(34)$ and (123) and swaps the two classes of 5-cycles, so this character is χ_Z. In particular, this automorphism is not conjugation by an element of A_5, which would preserve every conjugacy class.

NOTE Without the fact that the characters of A_5 are real, the computation above leaves two possibilities, because the dual Y^* is again a 3-dimensional irreducible representation, so it is Y or Z. Either $Y^* \cong Y$ and $Z^* \cong Z$, the characters are real, and $a = \pm\sqrt{5}$; or $Y^* \cong Z$, so that $\chi_Z = \overline{\chi_Y}$, and $a = \pm i\sqrt{5}$. Geometry also decides between them. The group A_5 is the rotation group of the regular icosahedron (Theorem 11.3.7), and the inclusion $A_5 \hookrightarrow SO(3) \subset GL_3(\mathbb{R}) \subset GL_3(\mathbb{C})$ is a 3-dimensional representation of A_5 by real matrices, so its character is real. Since the irreducible representations of A_5 have dimensions $1, 3, 3, 4, 5$, a 3-dimensional representation is $U^{\oplus 3}$, Y or Z, and this one is nontrivial, so it is Y or Z. Its character confirms the table: a rotation by the angle t has trace $1 + 2\cos t$, and rotations by $2\pi/5$, $4\pi/5$, $2\pi/3$ and π have traces $\frac{1+\sqrt{5}}{2}$, $\frac{1-\sqrt{5}}{2}$, 0 and -1.

The irrational character values have a geometric consequence: there is no regular icosahedron (or dodecahedron) in $\mathbb{R}^3$ all of whose vertices have rational coordinates. If there were, we could translate it so that its center, the average of its vertices, is at the origin, keeping the coordinates rational. Then in a basis consisting of three linearly independent vertices, each rotation of the icosahedron, which permutes the vertices, would have a matrix with rational entries, and all character values would be rational. But this representation is Y or Z, and both characters take the irrational values $\frac{1\pm\sqrt{5}}{2}$ on 5-cycles. This holds in any coordinate system, not just an orthonormal one.

14.5 The Representation Ring

What algebraic structure do the representations of a fixed finite group G carry? On isomorphism classes of representations there are two operations, $\oplus$ and $\otimes$. Both are commutative and associative up to isomorphism, and $\otimes$ distributes over $\oplus$: $(X \oplus Y) \otimes Z \cong (X \otimes Z) \oplus (Y \otimes Z)$. (The isomorphisms of Propositions 9.8.8 and 9.9.3 that express these laws are given by formulas such as $(x, y) \otimes z \mapsto (x \otimes z, y \otimes z)$, so they commute with the actions of G.) This is almost a ring; what is missing is additive inverses, and we add them formally.

Definition 14.5.1 (Representation ring)

Let $\hat{R}$ be the free abelian group on the set of isomorphism classes $[X]$ of representations of G; its elements are finite formal sums $\sum_j c_j[X_j]$ with $c_j \in \mathbb{Z}$. The ***representation ring*** $R(G)$ is the quotient of $\hat{R}$ by the subgroup N generated by all elements $[X] + [Y] - [X \oplus Y]$. Its multiplication is defined on generators by $[X] \cdot [Y] = [X \otimes Y]$ and extended bilinearly.

The multiplication is well defined on $\hat{R}$, which is free on the classes $[X]$, and it descends to $R(G) = \hat{R}/N$ because N is an ideal:

$$\big([X] + [Y] - [X \oplus Y]\big)[Z] = [X \otimes Z] + [Y \otimes Z] - [(X \otimes Z) \oplus (Y \otimes Z)] \in N,$$

using $(X \oplus Y) \otimes Z \cong (X \otimes Z) \oplus (Y \otimes Z)$, and similarly for products on the other side. It is commutative, associative and distributive because $\otimes$ is, and its identity is the class $[U]$ of the trivial representation, since $U \otimes X \cong X$. In $R(G)$ we have $[X] + [Y] = [X \oplus Y]$, and now differences $[X] - [Y]$ also make sense.

Proposition 14.5.2 (The character map)

The map $[X] \mapsto \chi_X$ induces an injective ring homomorphism

$$\Psi : R(G) \to \mathbb{C}_{\text{class}}(G), \qquad \sum_i a_i[V_i] \mapsto \sum_i a_i \chi_i,$$

where $\mathbb{C}_{\text{class}}(G)$ is a ring under pointwise addition and multiplication. The classes $[V_1], \dots, [V_k]$ of the irreducible representations form a basis of the abelian group $R(G)$, so $(R(G), +) \cong \mathbb{Z}^k$, where k is the number of conjugacy classes of G.

Proof. The homomorphism $\hat{R} \to \mathbb{C}_{\text{class}}(G)$ sending each generator $[X]$ to χ_X vanishes on N because $\chi_{X \oplus Y} = \chi_X + \chi_Y$, so it induces an additive map Ψ on $R(G)$. It is multiplicative because $\chi_{X \otimes Y} = \chi_X \chi_Y$, and $\Psi([U]) = 1$. By complete

reducibility, every representation satisfies $[X] = \sum_i a_i[V_i]$ in $R(G)$ with $a_i \geqslant 0$, so the classes $[V_i]$ generate $R(G)$ as an abelian group. If $\sum_i a_i[V_i] = 0$, then applying Ψ gives $\sum_i a_i\chi_i = 0$, so all $a_i = 0$ because the χ_i are linearly independent. Hence the $[V_i]$ form a basis, and Ψ is injective. Finally, k is the number of conjugacy classes by Theorem 13.6.2. □

General elements $\sum_i a_i[V_i]$ of $R(G)$ are called ***virtual representations***. The actual representations, those with all $a_i \geqslant 0$, form a cone inside $R(G) \cong \mathbb{Z}^k$, that is, a subset closed under addition. The image of Ψ is the lattice of ***virtual characters*** $\Lambda = \{\sum_i a_i\chi_i : a_i \in \mathbb{Z}\} \subset \mathbb{C}_{\text{class}}(G)$.

Example 14.5.3 (The representation ring of S_3)

By Example 13.3.6, $V \otimes V \cong U \oplus U' \oplus V$, and comparing characters gives $U' \otimes U' \cong U$ and $U' \otimes V \cong V$. So in $R(S_3)$, which has basis $[U]$, $[U']$, $[V]$ and identity $[U]$,

$$[U']^2 = [U], \qquad [U'][V] = [V], \qquad [V]^2 = [U] + [U'] + [V].$$

What, then, is an arbitrary class function the character of? Allowing complex coefficients, let $R(G)_{\mathbb{C}}$ be the complex vector space with basis $[V_1], \ldots, [V_k]$, with the multiplication of $R(G)$ extended bilinearly. The map Ψ extends to a $\mathbb{C}$-linear map $\Psi_{\mathbb{C}} : R(G)_{\mathbb{C}} \to \mathbb{C}_{\text{class}}(G)$, $\sum_i a_i[V_i] \mapsto \sum_i a_i\chi_i$ with $a_i \in \mathbb{C}$, which is still multiplicative.

Corollary 14.5.4 (Class functions as virtual characters)

The map $\Psi_{\mathbb{C}} : R(G)_{\mathbb{C}} \to \mathbb{C}_{\text{class}}(G)$ is an isomorphism of $\mathbb{C}$-algebras.

Proof. It is a multiplicative linear map that sends the basis $[V_1], \ldots, [V_k]$ of $R(G)_{\mathbb{C}}$ to the basis $\chi_1, \ldots, \chi_k$ of $\mathbb{C}_{\text{class}}(G)$ (Theorem 13.6.2). □

NOTE The theorems of Artin and Brauer (Theorems 15.3.4 and 15.3.5) describe the lattice Λ of virtual characters inside $\mathbb{C}_{\text{class}}(G)$ in terms of characters induced from subgroups of G.

Exercise 14.5.5 (Character Tables and the Representation Ring)

(1) Using the character table of S_4, show that $V \otimes V' \cong U' \oplus V \oplus V' \oplus W$.

(2) Let V be the standard representation of S_5. Show that $\bigwedge^3 V \cong V'$ by computing $\chi_{\bigwedge^3 V}$ from the formula

$$\chi_{\bigwedge^3 V}(g) = \tfrac{1}{6}\big(\chi_V(g)^3 - 3\chi_V(g)\chi_V(g^2) + 2\chi_V(g^3)\big),$$

and justify the formula.

(3) Verify column orthogonality (Proposition 13.6.5) for the character table of A_5.

(4) Using the character table of S_5, decompose $\bigwedge^2 V \otimes U'$ and $V \otimes W$ into irreducibles.

(5) Let $V_1, \dots, V_k$ be the irreducible representations of G, and write $[V_i][V_j] = \sum_\ell n_{ij}^\ell [V_\ell]$ in $R(G)$. Show that the structure constants n_{ij}^ℓ are nonnegative integers, that $n_{ij}^\ell = n_{ji}^\ell$, and that $n_{ij}^\ell = \frac{1}{|G|}\sum_{g\in G} \chi_i(g)\,\chi_j(g)\,\overline{\chi_\ell(g)}$.

(6) Let V be the standard representation of S_n, $n \geqslant 2$. Show that $\bigwedge^{n-1} V \cong U'$.

Solution. **(1)** The character of $V \otimes V'$ is $\chi_V\chi_{V'} = (9,-1,1,0,-1)$, in the order of the table. With the class sizes $1,6,3,8,6$,

$$\begin{aligned}
H(\chi_U, \chi_V\chi_{V'}) &= \tfrac{1}{24}(9-6+3+0-6) = 0,\\
H(\chi_{U'}, \chi_V\chi_{V'}) &= \tfrac{1}{24}(9+6+3+0+6) = 1,\\
H(\chi_W, \chi_V\chi_{V'}) &= \tfrac{1}{24}(18+0+6+0+0) = 1,\\
H(\chi_V, \chi_V\chi_{V'}) &= \tfrac{1}{24}(27-6-3+0+6) = 1,\\
H(\chi_{V'}, \chi_V\chi_{V'}) &= \tfrac{1}{24}(27+6-3+0-6) = 1.
\end{aligned}$$

So $V \otimes V' \cong U' \oplus V \oplus V' \oplus W$, and the dimensions agree: $1+3+3+2 = 9$. Alternatively, $V \otimes V' = (V \otimes V) \otimes U' \cong (U \oplus V \oplus V' \oplus W) \otimes U'$, and $W \otimes U' \cong W$.

(2) If g has eigenvalues $\lambda_1, \dots, \lambda_4$ on V with eigenvectors $v_1, \dots, v_4$, the vectors $v_i \wedge v_j \wedge v_l$, $i < j < l$, form a basis of $\bigwedge^3 V$ of eigenvectors with eigenvalues $\lambda_i\lambda_j\lambda_l$ (Proposition 9.10.19). Write $p_r = \sum_i \lambda_i^r = \chi_V(g^r)$ and $e_3 = \sum_{i<j<l} \lambda_i\lambda_j\lambda_l = \chi_{\bigwedge^3 V}(g)$. Expanding,

$$\begin{aligned}
p_1^3 &= p_3 + 3\sum_{i\neq j} \lambda_i^2\lambda_j + 6e_3,\\
p_1p_2 &= p_3 + \sum_{i\neq j} \lambda_i^2\lambda_j,
\end{aligned}$$

so $p_1^3 - 3p_1p_2 = -2p_3 + 6e_3$, which is the formula. Next we need $\chi_V(g^3)$. Cubing the representatives: $(12)^3 = (12)$, $((12)(34))^3 = (12)(34)$, $(123)^3 = e$, $((123)(45))^3 = (45)$, $(1234)^3 = (1432)$ and $(12345)^3 = (14253)$. So, in the order of the table,

$$\begin{aligned}\chi_V(g) &= (4,2,0,1,-1,0,-1),\\ \chi_V(g^2) &= (4,4,4,1,1,0,-1),\\ \chi_V(g^3) &= (4,2,0,4,2,0,-1).\end{aligned}$$

Entry by entry,

$$\begin{aligned}\chi_{\bigwedge^3 V} &= \tfrac{1}{6}\big(64-48+8,\ 8-24+4,\ 0,\ 1-3+8,\ -1+3+4,\ 0,\ -1-3-2\big)\\ &= (4,-2,0,1,1,0,-1),\end{aligned}$$

which is $\chi_{V'}$. So $\bigwedge^3 V \cong V'$.
(3) Write $\varphi = \frac{1+\sqrt{5}}{2}$ and $\bar\varphi = \frac{1-\sqrt{5}}{2}$, so that $\varphi + \bar\varphi = 1$ and $\varphi\bar\varphi = -1$; all entries are real. Using the rows U, Y, Z, V, W, the sums $\sum_i \chi_i(g_s)\chi_i(g_t)$ for $s = t$ are

$$e:\ 1+9+9+16+25 = 60, \quad (12)(34):\ 1+1+1+0+1 = 4,$$
$$(123):\ 1+0+0+1+1 = 3, \qquad 5\text{-cycles}:\ 1+\varphi^2+\bar\varphi^2+1+0 = 5,$$

using $\varphi^2 + \bar\varphi^2 = (\varphi+\bar\varphi)^2 - 2\varphi\bar\varphi = 3$ for either class of 5-cycles. These are $60/1$, $60/15$, $60/20$ and $60/12$. For $s \neq t$:

$$\begin{aligned} e,\ (12)(34):&\ 1-3-3+0+5 = 0,\\ e,\ (123):&\ 1+0+0+4-5 = 0,\\ e,\ 5\text{-cycles}:&\ 1+3(\varphi+\bar\varphi)-4+0 = 0,\\ (12)(34),\ (123):&\ 1+0+0+0-1 = 0,\\ (12)(34),\ 5\text{-cycles}:&\ 1-(\varphi+\bar\varphi)+0+0 = 0,\\ (123),\ 5\text{-cycles}:&\ 1+0+0-1+0 = 0,\end{aligned}$$

and for the two classes of 5-cycles, $1 + 2\varphi\bar\varphi + 1 + 0 = 0$.
(4) The character of $\bigwedge^2 V \otimes U'$ is $\chi_{\bigwedge^2 V}\,\mathrm{sgn}$. The odd classes are those of (12), $(123)(45)$ and (1234), and $\chi_{\bigwedge^2 V}$ vanishes on all three, so $\chi_{\bigwedge^2 V}\,\mathrm{sgn} = \chi_{\bigwedge^2 V}$ and $\bigwedge^2 V \otimes U' \cong \bigwedge^2 V$.
The character of $V \otimes W$ is $\chi_V\chi_W = (20,2,0,-1,-1,0,0)$. With the class sizes $1, 10, 15, 20, 20, 30, 24$,

$$H(\chi_U, \chi_V\chi_W) = \tfrac{1}{120}(20+20+0-20-20) = 0,$$

$$\begin{aligned}
H(\chi_{U'}, \chi_V\chi_W) &= \tfrac{1}{120}(20 - 20 + 0 - 20 + 20) = 0,\\
H(\chi_V, \chi_V\chi_W) &= \tfrac{1}{120}(80 + 40 + 0 - 20 + 20) = 1,\\
H(\chi_{V'}, \chi_V\chi_W) &= \tfrac{1}{120}(80 - 40 + 0 - 20 - 20) = 0,\\
H(\chi_W, \chi_V\chi_W) &= \tfrac{1}{120}(100 + 20 + 0 + 20 - 20) = 1,\\
H(\chi_{W'}, \chi_V\chi_W) &= \tfrac{1}{120}(100 - 20 + 0 + 20 + 20) = 1,\\
H(\chi_{\bigwedge^2 V}, \chi_V\chi_W) &= \tfrac{1}{120}(120 + 0 + 0 + 0 + 0) = 1,
\end{aligned}$$

where we have omitted the last two classes, on which $\chi_V\chi_W$ vanishes. So $V \otimes W \cong V \oplus W \oplus W' \oplus \bigwedge^2 V$, of dimension $4 + 5 + 5 + 6 = 20$.

(5) We have $[V_i][V_j] = [V_i \otimes V_j]$, and by Corollary 13.5.6, $V_i \otimes V_j \cong \bigoplus_\ell V_\ell^{\oplus m_\ell}$ with $m_\ell = H(\chi_\ell, \chi_i\chi_j)$, a nonnegative integer. Since the $[V_\ell]$ form a basis of $R(G)$, $n_{ij}^\ell = m_\ell = H(\chi_\ell, \chi_i\chi_j) = \frac{1}{|G|}\sum_g \overline{\chi_\ell(g)}\,\chi_i(g)\,\chi_j(g)$, which is symmetric in i and j.

(6) Since $\dim V = n - 1$, the space $\bigwedge^{n-1} V$ is 1-dimensional, and $\sigma \in S_n$ acts on it by $\det(\sigma|_V)$ (Theorem 9.11.3). The permutation representation is $\mathbb{C}^n = U \oplus V$, so the permutation matrix of σ is conjugate to a block-diagonal matrix with blocks 1 and the matrix of $\sigma|_V$, and $\det(\sigma|_V)$ equals the determinant of the permutation matrix, which is $\operatorname{sgn}\sigma$ (Proposition 5.4.24). So σ acts on $\bigwedge^{n-1} V$ as multiplication by $\operatorname{sgn}\sigma$, and $\bigwedge^{n-1} V \cong U'$.

CHAPTER 15

Restriction and Induction

Throughout this chapter G is a finite group, $H \leqslant G$ is a subgroup, and representations are finite-dimensional complex representations. A representation of G gives a representation of H by forgetting part of the action. Induction goes the other way: it builds a representation of G out of a representation of H. Frobenius reciprocity relates the two constructions, and it lets us transport information about the representations of one group to the other.

Since the letter H now denotes a subgroup, we write the Hermitian inner product on class functions of Definition 13.5.3 as

$$\langle \alpha, \beta \rangle_G = \frac{1}{|G|} \sum_{g \in G} \overline{\alpha(g)}\, \beta(g),$$

conjugate-linear in the first variable as before, and $\langle \cdot, \cdot \rangle_H$ for the same inner product on class functions of H. Recall that $\dim \mathrm{Hom}_G(V, W) = \langle \chi_V, \chi_W \rangle_G$ for representations V, W of G (Theorem 13.5.4); in particular, the multiplicity of an irreducible representation V in W is $\langle \chi_V, \chi_W \rangle_G$ (Corollary 13.5.6), and W is irreducible if and only if $\langle \chi_W, \chi_W \rangle_G = 1$ (Corollary 13.5.7).

15.1 Restriction

Definition 15.1.1 (Restriction)

Let V be a representation of G, given by $\rho : G \to \mathrm{GL}(V)$. The ***restriction*** of V to H, written $\mathrm{Res}^G_H V$ (or $\mathrm{Res}\, V$ when G and H are clear), is the representation $\rho|_H : H \to \mathrm{GL}(V)$ of H on the same space: we forget the action of the elements outside H.

A G-equivariant linear map is in particular H-equivariant. So restriction is a functor (Definition 8.2.1)

$$\mathrm{Res}^G_H : \mathrm{Rep}(G) \to \mathrm{Rep}(H),$$

where $\mathrm{Rep}(G)$ is the category whose objects are the representations of G and whose morphisms are the G-equivariant linear maps, and similarly for $\mathrm{Rep}(H)$. The character of $\mathrm{Res}^G_H V$ is the restriction $\chi_V|_H$ of the character of V.

An irreducible representation of G need not remain irreducible when restricted to H.

Example 15.1.2 (Restricting from S_3 to A_3)

Let V be the standard representation of S_3, and let $A_3 = \langle (123) \rangle \cong \mathbb{Z}/3$. Since A_3 is abelian, its irreducible representations are 1-dimensional (Proposition 13.1.7): they are $\chi_1, \chi_\omega, \chi_{\omega^2}$, where $\omega = e^{2\pi i/3}$ and χ_{ω^a} sends (123) to ω^a. The character of $\mathrm{Res}\, V$ takes the values $2, -1, -1$ at $e, (123), (132)$. Using $1 + \omega + \omega^2 = 0$,

$$\langle \chi_{\omega^a}, \chi_{\mathrm{Res}\, V} \rangle_{A_3} = \tfrac{1}{3}\left(2 - \omega^{-a} - \omega^{-2a}\right) = \begin{cases} 0 & \text{if } a = 0, \\ 1 & \text{if } a = 1, 2. \end{cases}$$

So $\mathrm{Res}^{S_3}_{A_3} V \cong \chi_\omega \oplus \chi_{\omega^2}$.

The rest of the chapter asks whether we can go in the opposite direction, from representations of H to representations of G.

15.2 Induced Representations

Suppose V is a representation of G and $W \subset V$ is a subspace invariant under H, so that W is a subrepresentation of $\mathrm{Res}^G_H V$, but not necessarily invariant under all of G. For $g \in G$, the subspace $gW \subset V$ depends only on the coset gH, since $ghW = g(hW) = gW$ for $h \in H$. Moreover, gW is invariant under the conjugate subgroup gHg^{-1}: for $h \in H$ and $w \in W$,

$$(ghg^{-1})(gw) = g(hw) \in gW.$$

In other words, conjugation $c_g : H \to gHg^{-1}$, $h \mapsto ghg^{-1}$, and the linear isomorphism $W \to gW$, $w \mapsto gw$, carry the action of H on W to the action of gHg^{-1} on gW. With ρ and ρ' denoting these two actions, the square

$$\begin{array}{ccc} H & \xrightarrow{\ \rho\ } & GL(W) \\ {\scriptstyle c_g}\downarrow{\scriptstyle \wr} & & \downarrow{\scriptstyle A \mapsto gAg^{-1}} \\ gHg^{-1} & \xrightarrow{\ \rho'\ } & GL(gW) \end{array}$$

commutes.

The simplest situation is when V is the direct sum of these translates,

$$V = \bigoplus_{\sigma \in G/H} \sigma W,$$

where σ runs over a set of representatives of the cosets of H. Then $\dim V = [G : H] \cdot \dim W$, and, as we show below, the representation W of H determines the whole representation V of G. In general there is no reason for this to happen; when it does, we say that V is induced from W.

Definition 15.2.1 (Induced representation)

Let W be a representation of H. A representation V of G, together with an H-invariant subspace $W \subset V$ (so that W is a subrepresentation of $\operatorname{Res}^G_H V$), is ***induced by*** W if, as a vector space,

$$V = \bigoplus_{\sigma \in G/H} \sigma W.$$

We then write $V = \operatorname{Ind}^G_H W$, or $\operatorname{Ind} W$.

Concretely, fix representatives $\sigma_1, \dots, \sigma_k$ of the cosets of H, where $k = [G : H]$; we always take $\sigma_1 = e$. Every $v \in V$ can be written uniquely as $v = \sigma_1 w_1 + \cdots + \sigma_k w_k$ with $w_i \in W$. For $g \in G$ and each i, there are a unique index j and a unique $h \in H$ with $g\sigma_i = \sigma_j h$: the coset $\sigma_j H$ is the one containing $g\sigma_i$, and $h = \sigma_j^{-1} g \sigma_i$. Then

$$g(\sigma_i w) = (g\sigma_i)w = (\sigma_j h)w = \sigma_j(hw). \tag{15.1}$$

So g permutes the summands $\sigma_i W$ in the same way as it permutes the cosets in G/H, and it maps $\sigma_i W$ to $\sigma_j W$ through the action of an element of H on W.

Theorem 15.2.2 (Existence and uniqueness of induced representations)

Let W be a representation of H. There is a representation of G induced by W, and it is unique up to isomorphism: if V and V' are both induced by W, there is an isomorphism of representations $V \to V'$ that is the identity on W.

Proof. Uniqueness. Let V and V' be induced by W, and fix coset representatives $\sigma_1 = e, \sigma_2, \dots, \sigma_k$. Define a linear map $\Phi : V \to V'$ by $\Phi(\sigma_i w) = \sigma_i w$ for $w \in W$, where on the left σ_i acts in V and on the right in V'. It is a linear isomorphism, because V and V' are the direct sums of the subspaces $\sigma_i W$, each isomorphic to

W via $w \mapsto \sigma_i w$. If $g\sigma_i = \sigma_j h$ with $h \in H$, then (15.1) holds in both V and V', so

$$\Phi\big(g(\sigma_i w)\big) = \Phi\big(\sigma_j(hw)\big) = \sigma_j(hw) = g(\sigma_i w) = g\,\Phi(\sigma_i w).$$

Hence Φ is G-equivariant, and it is the identity on $W = \sigma_1 W$.

Existence. Let $V = W^{\oplus k}$, and write $\sigma_i w$ for the vector with w in the ith place and 0 elsewhere, so that $V = \bigoplus_i \sigma_i W$ with σ_i a formal symbol. For $g \in G$, define a linear map $\rho(g) : V \to V$ summand by summand, following (15.1): $\rho(g)(\sigma_i w) = \sigma_j(hw)$, where $g\sigma_i = \sigma_j h$. Then $\rho(e) = \mathrm{id}$, since $e\sigma_i = \sigma_i e$. If $g\sigma_i = \sigma_j h$ and $g'\sigma_j = \sigma_l h'$, then $g'g\sigma_i = g'\sigma_j h = \sigma_l(h'h)$, so

$$\rho(g'g)(\sigma_i w) = \sigma_l(h'hw) = \rho(g')\big(\sigma_j(hw)\big) = \rho(g')\rho(g)(\sigma_i w).$$

Thus $\rho(g'g) = \rho(g')\rho(g)$; in particular $\rho(g)$ is invertible with inverse $\rho(g^{-1})$, and $\rho : G \to \mathrm{GL}(V)$ is a representation. The subspace $W = \sigma_1 W$ is H-invariant: for $h \in H$ we have $h\sigma_1 = \sigma_1 h$, so $\rho(h)(\sigma_1 w) = \sigma_1(hw)$, and H acts on it as on W. Finally $\rho(\sigma_i)(\sigma_1 w) = \sigma_i w$, because $\sigma_i\sigma_1 = \sigma_i e$; so the summand $\sigma_i W$ is the translate of W by σ_i, and $V = \bigoplus_i \sigma_i W$ is induced by W. □

The construction in the proof depends on the choice of the σ_i, but by uniqueness different choices give isomorphic representations. In every case

$$\dim \mathrm{Ind}_H^G W = [G : H] \cdot \dim W.$$

15.2.1 Examples

We write U_G and U_H for the trivial 1-dimensional representations of G and H.

Example 15.2.3 (Permutation representations are induced)

Let $\mathbb{C}[G/H]$ be the permutation representation of G on the set G/H of left cosets: it has basis $\{e_{\sigma H}\}_{\sigma H \in G/H}$, and $g \cdot e_{\sigma H} = e_{g\sigma H}$. The vector e_H is fixed by H, so $W = \mathrm{span}(e_H)$ is an H-invariant line on which H acts trivially, and $gW = \mathrm{span}(e_{gH})$. Hence $\mathbb{C}[G/H] = \bigoplus_{gH \in G/H} gW$, that is,

$$\mathbb{C}[G/H] \cong \mathrm{Ind}_H^G U_H.$$

Example 15.2.4 (The regular representation is induced)

Let $R_G = \mathbb{C}[G]$ be the regular representation of G, with basis $\{e_g\}_{g \in G}$, and let $W = \mathrm{span}\{e_h : h \in H\}$. Then W is H-invariant and isomorphic to the regular representation R_H of H, and $\sigma W = \mathrm{span}\{e_g : g \in \sigma H\}$. Since G is the disjoint

union of the cosets of H, we get $\mathbb{C}[G] = \bigoplus_{\sigma \in G/H} \sigma W$, so $R_G \cong \operatorname{Ind}_H^G R_H$. Taking $H = \{e\}$ gives $R_G \cong \operatorname{Ind}_{\{e\}}^G \mathbb{C}$.

15.2.2 Properties of Induction

Proposition 15.2.5 (Basic properties of induction)

Let W, W' be representations of H and V a representation of G.

(i) $\operatorname{Ind}_H^G(W \oplus W') \cong \operatorname{Ind}_H^G W \oplus \operatorname{Ind}_H^G W'$.

(ii) $\operatorname{Ind}_H^G\big(\operatorname{Res}_H^G V \otimes W\big) \cong V \otimes \operatorname{Ind}_H^G W$.

(iii) $\operatorname{Ind}_H^G \operatorname{Res}_H^G V \cong V \otimes \mathbb{C}[G/H]$, where $\mathbb{C}[G/H]$ is the permutation representation on G/H.

Proof. (i) In $X = \operatorname{Ind} W \oplus \operatorname{Ind} W'$, the subspace $W \oplus W'$ is H-invariant, and $\sigma(W \oplus W') = \sigma W \oplus \sigma W'$. So $X = \bigoplus_{\sigma \in G/H} \sigma(W \oplus W')$ is induced by $W \oplus W'$, and Theorem 15.2.2 gives the isomorphism.

(ii) Write $\operatorname{Ind} W = \bigoplus_{\sigma \in G/H} \sigma W$. Tensor products distribute over direct sums, so

$$V \otimes \operatorname{Ind} W = \bigoplus_{\sigma \in G/H} (V \otimes \sigma W).$$

Since $\sigma(u \otimes w) = \sigma u \otimes \sigma w$ and σ acts bijectively on V, we have $V \otimes \sigma W = \sigma(V \otimes W)$. The subspace $V \otimes W$ is H-invariant, and as a representation of H it is $\operatorname{Res} V \otimes W$. So $V \otimes \operatorname{Ind} W$ is induced by $\operatorname{Res} V \otimes W$.

(iii) Take $W = U_H$ in (ii): $\operatorname{Res} V \otimes U_H \cong \operatorname{Res} V$, and $\operatorname{Ind} U_H \cong \mathbb{C}[G/H]$ by Example 15.2.3. □

NOTE Induction does not commute with tensor products. If $H \neq G$ and $W, W' \neq 0$, then $\dim \operatorname{Ind}(W \otimes W') = [G : H] \dim W \dim W'$, while $\dim(\operatorname{Ind} W \otimes \operatorname{Ind} W') = [G : H]^2 \dim W \dim W'$, so $\operatorname{Ind}(W \otimes W') \ncong \operatorname{Ind} W \otimes \operatorname{Ind} W'$.

15.2.3 The Character of an Induced Representation

Proposition 15.2.6 (Character formula for induction)

Let W be a representation of H, and let $\sigma_1, \dots, \sigma_k$ be representatives of the cosets of H in G. For $g \in G$,

$$\chi_{\operatorname{Ind} W}(g) = \sum_{\substack{1 \leqslant i \leqslant k \\ \sigma_i^{-1} g \sigma_i \in H}} \chi_W(\sigma_i^{-1} g \sigma_i) = \frac{1}{|H|} \sum_{\substack{s \in G \\ s^{-1} g s \in H}} \chi_W(s^{-1} g s).$$

Proof. By (15.1), g maps $\sigma_i W$ to $\sigma_j W$, where $g\sigma_i \in \sigma_j H$. In a basis of $\operatorname{Ind} W$ made of bases of the summands $\sigma_i W$, the matrix of g is a block matrix, and the blocks with $j \neq i$ lie off the diagonal, so they contribute nothing to the trace. If $j = i$, then $h = \sigma_i^{-1} g \sigma_i \in H$ and $g(\sigma_i w) = \sigma_i(hw)$. So under the isomorphism $W \to \sigma_i W$, $w \mapsto \sigma_i w$, the restriction of g to $\sigma_i W$ corresponds to h acting on W, and its trace is $\chi_W(h)$. Since $g\sigma_i \in \sigma_i H$ if and only if $\sigma_i^{-1} g \sigma_i \in H$, summing over i gives the first formula.

For the second, every $s \in G$ can be written uniquely as $s = \sigma_i h'$ with $h' \in H$, and then $s^{-1} g s = h'^{-1}(\sigma_i^{-1} g \sigma_i) h'$. This lies in H if and only if $\sigma_i^{-1} g \sigma_i$ does, and in that case $\chi_W(s^{-1} g s) = \chi_W(\sigma_i^{-1} g \sigma_i)$, because χ_W is a class function on H. So each coset $\sigma_i H$ contributes either nothing or $|H|$ copies of the term $\chi_W(\sigma_i^{-1} g \sigma_i)$ to the second sum. □

In particular $\chi_{\operatorname{Ind} W}(e) = [G : H] \dim W$, and $\chi_{\operatorname{Ind} W}(g) = 0$ whenever g is not conjugate in G to an element of H.

Example 15.2.7 (Inducing from A_3 to S_3)

With the notation of Example 15.1.2, take $G = S_3$, $H = A_3$, and coset representatives $\sigma_1 = e$, $\sigma_2 = (12)$. Let us compute $\chi = \chi_{\operatorname{Ind} \chi_\omega}$.

- $\chi(e) = [S_3 : A_3] \cdot 1 = 2$.
- A transposition is not conjugate to any element of A_3, so $\chi((12)) = 0$.
- For $g = (123)$, both conjugates $\sigma_1^{-1} g \sigma_1 = (123)$ and $\sigma_2^{-1} g \sigma_2 = (12)(123)(12) = (132)$ lie in A_3, so $\chi((123)) = \omega + \omega^2 = -1$.

So $\chi = (2, 0, -1)$ on the classes of e, (12), (123). This is the character of the standard representation V of S_3, and since a representation is determined by its character (Corollary 13.5.6), $\operatorname{Ind}_{A_3}^{S_3} \chi_\omega \cong V$.

15.3 Frobenius Reciprocity

The key property of induction is that it is adjoint to restriction.

Theorem 15.3.1 (Frobenius reciprocity)
Let V be a representation of G and W a representation of $H \leqslant G$. Every H-equivariant linear map $\varphi : W \to \operatorname{Res}_H^G V$ extends uniquely to a G-equivariant linear map $\tilde{\varphi} : \operatorname{Ind}_H^G W \to V$. That is, restriction to $W \subset \operatorname{Ind}_H^G W$ is an isomorphism of vector spaces

$$\operatorname{Hom}_G(\operatorname{Ind}_H^G W, V) \xrightarrow{\sim} \operatorname{Hom}_H(W, \operatorname{Res}_H^G V).$$

Proof. Choose coset representatives $\sigma_1 = e, \sigma_2, \dots, \sigma_k$, so that $\operatorname{Ind} W = \bigoplus_{i=1}^k \sigma_i W$ and $W = \sigma_1 W$.

Restriction is well defined. If $\tilde{\varphi} : \operatorname{Ind} W \to V$ is G-equivariant, it is in particular H-equivariant, so its restriction to the H-invariant subspace W lies in $\operatorname{Hom}_H(W, \operatorname{Res} V)$. Restriction is clearly linear.

Uniqueness of the extension. Let $\varphi \in \operatorname{Hom}_H(W, \operatorname{Res} V)$. If $\tilde{\varphi}$ is G-equivariant and $\tilde{\varphi}|_W = \varphi$, then for $w \in W$,

$$\tilde{\varphi}(\sigma_i w) = \sigma_i\, \tilde{\varphi}(w) = \sigma_i\, \varphi(w),$$

where on the right σ_i acts on $\varphi(w) \in V$. This determines $\tilde{\varphi}$ on each summand $\sigma_i W$, hence on $\operatorname{Ind} W$.

Existence of the extension. Define $\tilde{\varphi}$ on each summand by $\tilde{\varphi}(\sigma_i w) = \sigma_i\, \varphi(w)$; it is linear, and $\tilde{\varphi}|_W = \varphi$ since $\sigma_1 = e$. Let $g \in G$ and $g\sigma_i = \sigma_j h$ with $h \in H$. By (15.1),

$$\tilde{\varphi}\big(g(\sigma_i w)\big) = \tilde{\varphi}\big(\sigma_j(hw)\big) = \sigma_j\, \varphi(hw) = \sigma_j h\, \varphi(w) = g\sigma_i\, \varphi(w) = g\, \tilde{\varphi}(\sigma_i w),$$

using the H-equivariance of φ in the third equality and $\sigma_j h = g\sigma_i$ in the fourth. So $\tilde{\varphi}$ is G-equivariant on each summand, hence on $\operatorname{Ind} W$.

Thus $\varphi \mapsto \tilde{\varphi}$ is inverse to restriction, and restriction is an isomorphism. □

Taking dimensions on both sides, and using $\dim \operatorname{Hom}_G(X, Y) = \langle \chi_X, \chi_Y \rangle_G$, we obtain the character form.

Corollary 15.3.2 (Frobenius reciprocity for characters)
For a representation V of G and a representation W of H,

$$\langle \chi_{\operatorname{Ind}_H^G W}, \chi_V \rangle_G = \langle \chi_W, \chi_{\operatorname{Res}_H^G V} \rangle_H.$$

Both sides are dimensions, hence real, so we may also exchange the two arguments on each side. The character form can also be proved by a direct calculation from the formula for $\chi_{\operatorname{Ind} W}$; see Exercise 15.5.2(4).

NOTE If V is an irreducible representation of G and W an irreducible representation of H, Corollary 15.3.2 says: the number of times W appears in $\operatorname{Res}^G_H V$ equals the number of times V appears in $\operatorname{Ind}^G_H W$.

Example 15.3.3 ($\operatorname{Ind}^{S_3}_{A_3} \chi_\omega$ by reciprocity)
By Example 15.1.2, $\operatorname{Res} V = \chi_\omega \oplus \chi_{\omega^2}$, while the trivial and sign representations U and U' of S_3 both restrict to χ_1, since A_3 consists of even permutations. So V appears once in $\operatorname{Ind} \chi_\omega$, and U and U' do not appear. As $\dim \operatorname{Ind} \chi_\omega = 2 = \dim V$, we recover $\operatorname{Ind}^{S_3}_{A_3} \chi_\omega \cong V$.

15.3.1 The Theorems of Artin and Brauer

Much of the interest of induced representations comes from two deep theorems, which say that all characters can be built from characters induced from very simple subgroups. We state them without proof; see Serre, *Linear Representations of Finite Groups*, Chapters 9 and 10.

Theorem 15.3.4 (Artin)
Every character of a representation of G is a linear combination, with rational coefficients, of characters of representations induced from cyclic subgroups of G.

A finite group is ***elementary*** if it is isomorphic to a product $C \times P$, where P is a p-group for some prime p and C is cyclic of order prime to p.

Theorem 15.3.5 (Brauer)
Every character of a representation of G is a linear combination, with integer coefficients, of characters of representations induced from 1-dimensional representations of elementary subgroups of G.

15.4 Irreducibility of Induced Representations

An induced representation need not be irreducible. For instance, $\operatorname{Ind}^G_{\{e\}} \mathbb{C} = R_G$ contains every irreducible representation of G. Since $\operatorname{Ind} W$ is irreducible if and only if $\langle \chi_{\operatorname{Ind} W}, \chi_{\operatorname{Ind} W} \rangle_G = 1$, and Frobenius reciprocity gives

$$\langle \chi_{\operatorname{Ind} W}, \chi_{\operatorname{Ind} W} \rangle_G = \langle \chi_W, \chi_{\operatorname{Res} \operatorname{Ind} W} \rangle_H,$$

the question comes down to understanding $\operatorname{Res}_H^G \operatorname{Ind}_H^G W$. We do this when H is normal.

Let $H \trianglelefteq G$, let W be a representation of H given by $\rho : H \to \mathrm{GL}(W)$, and let $g \in G$. The ***conjugate representation*** W^g is the space W with $h \in H$ acting by $\rho(g^{-1}hg)$; this makes sense because $g^{-1}hg \in H$. Its character is $\chi_{W^g}(h) = \chi_W(g^{-1}hg)$. The representations W and W^g have the same invariant subspaces, so W^g is irreducible if W is. For $h_0 \in H$, the map $\rho(h_0) : W^{gh_0} \to W^g$ is an isomorphism of representations, since $\rho(h_0)\rho(h_0^{-1}g^{-1}hgh_0) = \rho(g^{-1}hg)\rho(h_0)$; so the isomorphism class of W^g depends only on the coset gH, and $W^{h_0} \cong W$.

Lemma 15.4.1 (Restricting an induced representation to a normal subgroup)
Let $H \trianglelefteq G$, let $\sigma_1, \dots, \sigma_k$ be representatives of the cosets of H, and let W be a representation of H. Then

$$\operatorname{Res}_H^G \operatorname{Ind}_H^G W \cong W^{\sigma_1} \oplus \cdots \oplus W^{\sigma_k}.$$

Proof. Each summand $\sigma_i W$ of $\operatorname{Ind} W$ is H-invariant: for $h \in H$, we have $h\sigma_i = \sigma_i(\sigma_i^{-1}h\sigma_i)$ with $\sigma_i^{-1}h\sigma_i \in H$, so $h(\sigma_i w) = \sigma_i\big((\sigma_i^{-1}h\sigma_i)w\big)$. This says exactly that the linear isomorphism $W \to \sigma_i W$, $w \mapsto \sigma_i w$, is an isomorphism of representations of H from W^{σ_i} to $\sigma_i W$. □

Proposition 15.4.2 (Induction from a normal subgroup)
Let $H \trianglelefteq G$ and let W be an irreducible representation of H. Then $\langle \chi_{\operatorname{Ind} W}, \chi_{\operatorname{Ind} W}\rangle_G$ is the number of cosets $gH \in G/H$ with $W^g \cong W$. In particular, $\operatorname{Ind}_H^G W$ is irreducible if and only if $W^g \not\cong W$ for every $g \in G \setminus H$.

Proof. By Frobenius reciprocity and Lemma 15.4.1,

$$\langle \chi_{\operatorname{Ind} W}, \chi_{\operatorname{Ind} W}\rangle_G = \langle \chi_W, \chi_{\operatorname{Res}\operatorname{Ind} W}\rangle_H = \sum_{i=1}^{k} \langle \chi_W, \chi_{W^{\sigma_i}}\rangle_H.$$

Each W^{σ_i} is irreducible, so the ith term is 1 if $W^{\sigma_i} \cong W$ and 0 otherwise. The coset H itself always counts, since $W^h \cong W$ for $h \in H$; so the sum equals 1 exactly when no other coset counts. □

NOTE For a subgroup H that is not normal, Mackey's analysis of $\operatorname{Res}_H^G \operatorname{Ind}_H^G W$ leads to the following criterion, which we state without proof (see Serre, *Linear Representations of Finite Groups*, §7.4). Let W be an irreducible representation of H, given by ρ, and for $g \in G$ let $H_g = H \cap gHg^{-1}$. Then $\operatorname{Ind}_H^G W$ is irreducible if and only if, for every $g \in G \setminus H$, the two representa-

tions $x \mapsto \rho(x)$ and $x \mapsto \rho(g^{-1}xg)$ of H_g have no irreducible constituent in common. When H is normal, $H_g = H$ and this is Proposition 15.4.2.

Example 15.4.3 (Inducing from a subgroup of index 2)

Let $[G : H] = 2$. Then H is normal, since its two left cosets H and $G \setminus H$ are also its two right cosets. Fix $g \in G \setminus H$ and an irreducible representation W of H. By Proposition 15.4.2:

- if $W^g \not\cong W$, then $\operatorname{Ind}_H^G W$ is irreducible;
- if $W^g \cong W$, then $\langle \chi_{\operatorname{Ind} W}, \chi_{\operatorname{Ind} W} \rangle_G = 2$, so $\operatorname{Ind} W \cong X \oplus Y$ with X and Y irreducible and not isomorphic.

In the second case $\dim X = \dim Y = \dim W$. Indeed, by Frobenius reciprocity $\langle \chi_W, \chi_{\operatorname{Res} X} \rangle_H = \langle \chi_{\operatorname{Ind} W}, \chi_X \rangle_G = 1$, so $\operatorname{Res} X$ contains W and $\dim X \geqslant \dim W$; likewise $\dim Y \geqslant \dim W$; and $\dim X + \dim Y = 2 \dim W$. For $G = S_3$ and $H = A_3$, take $g = (12)$. Since $(12)^{-1}(123)(12) = (132)$, we get $\chi_\omega^{(12)} = \chi_{\omega^2} \not\cong \chi_\omega$, so $\operatorname{Ind} \chi_\omega$ is irreducible, in agreement with Example 15.2.7. On the other hand the trivial representation χ_1 of A_3 is its own conjugate, and $\operatorname{Ind} \chi_1 = \mathbb{C}[S_3/A_3]$ splits as $U \oplus U'$ (Exercise 15.5.2(2)).

15.5 Example: $S_4 \supset S_3$

Example 15.5.1 (Restriction and induction between S_4 and S_3)

Let $G = S_4$, with irreducible representations U_4, U_4', V_4, V_4', W of dimensions $1, 1, 3, 3, 2$ (called U, U', V, V', W in the character table of S_4, Section 14.1), and let $H = S_3 \subset S_4$ be the stabilizer of 4, with irreducible representations U_3, U_3', V_3 of dimensions $1, 1, 2$. The conjugacy classes of S_3 are those of e, (12), (123). Evaluating the characters of S_4 at these elements gives the characters of the restrictions:

	e	(12)	(123)
$\operatorname{Res} U_4$	1	1	1
$\operatorname{Res} U_4'$	1	-1	1
$\operatorname{Res} V_4$	3	1	0
$\operatorname{Res} V_4'$	3	-1	0
$\operatorname{Res} W$	2	0	-1

	e	(12)	(123)
U_3	1	1	1
U_3'	1	-1	1
V_3	2	0	-1

Comparing with the character table of S_3 (on the right),

$$\operatorname{Res} U_4 = U_3, \qquad \operatorname{Res} U_4' = U_3', \qquad \operatorname{Res} W = V_3,$$

$$\operatorname{Res} V_4 = U_3 \oplus V_3, \qquad \operatorname{Res} V_4' = U_3' \oplus V_3.$$

These can also be seen directly. The permutation representation $\mathbb{C}^4 = U_4 \oplus V_4$ restricts to $\mathbb{C}^3 \oplus \mathbb{C}$, since S_3 permutes the first three coordinates and fixes the fourth; so $\operatorname{Res}(U_4 \oplus V_4) = U_3 \oplus V_3 \oplus U_3$. Next, $V_4' = V_4 \otimes U_4'$ and $V_3 \otimes U_3' \cong V_3$. Finally, W factors through the quotient of S_4 by $K = \{e, (12)(34), (13)(24), (14)(23)\}$, and S_3 maps isomorphically onto S_4/K (as $S_3 \cap K = \{e\}$ and both groups have order 6), so $\operatorname{Res} W$ is a 2-dimensional irreducible representation of S_3, which must be V_3.

By Frobenius reciprocity, the multiplicity of an irreducible X of S_4 in $\operatorname{Ind} Y$, for Y an irreducible of S_3, is the multiplicity of Y in $\operatorname{Res} X$. Reading the restrictions above in reverse:

- $\operatorname{Ind}_{S_3}^{S_4} V_3 = V_4 \oplus V_4' \oplus W$, of dimension $3 + 3 + 2 = 8 = [S_4 : S_3] \cdot 2$;
- $\operatorname{Ind}_{S_3}^{S_4} U_3 = U_4 \oplus V_4$, of dimension $1 + 3 = 4 = [S_4 : S_3] \cdot 1$; this is the permutation representation $\mathbb{C}^4 = \mathbb{C}[S_4/S_3]$, as Example 15.2.3 predicts;
- $\operatorname{Ind}_{S_3}^{S_4} U_3' = U_4' \oplus V_4'$, of dimension $1 + 3 = 4$.

Exercise 15.5.2 (Exercises)

(1) Let $H = \{e\}$, so that a representation of H is just a vector space W. Show that $\operatorname{Ind}_{\{e\}}^{G} W \cong W_0 \otimes R_G$, where W_0 is W with the trivial action of G and R_G is the regular representation. In other words, $\operatorname{Ind}_{\{e\}}^{G} W$ is a sum of $\dim W$ copies of R_G.

(2) Let $n \geqslant 2$. Show that $\operatorname{Ind}_{A_n}^{S_n} U_{A_n} \cong U \oplus U'$, the sum of the trivial and sign representations of S_n.

(3) Using Frobenius reciprocity and the character table of S_5, decompose $\operatorname{Ind}_{S_4}^{S_5} V_4$ into irreducibles, where V_4 is the standard representation of S_4 and $S_4 \subset S_5$ is the stabilizer of 5.

(4) Prove Corollary 15.3.2 directly from the formula for $\chi_{\operatorname{Ind} W}$ in Proposition 15.2.6, without using Theorem 15.3.1.

(5) (Transitivity of induction.) If $K \leqslant H \leqslant G$ and W is a representation of K, show that $\operatorname{Ind}_H^G \operatorname{Ind}_K^H W \cong \operatorname{Ind}_K^G W$.

Solution. **(1)** Apply Proposition 15.2.5(ii) with $V = W_0$ and the trivial representation $\mathbb{C}$ of $\{e\}$: $\operatorname{Ind}(\operatorname{Res} W_0 \otimes \mathbb{C}) \cong W_0 \otimes \operatorname{Ind} \mathbb{C}$, and $\operatorname{Ind}_{\{e\}}^{G} \mathbb{C} \cong R_G$ by Example 15.2.4, while $\operatorname{Res} W_0 \otimes \mathbb{C} = W$. Concretely, $\operatorname{Ind} W = \bigoplus_{g \in G} gW$, with g' mapping gw to $(g'g)w$. If $w_1, \ldots, w_m$ is a basis of W, then for each

l the vectors gw_l ($g \in G$) span a copy of R_G, and $\operatorname{Ind} W$ is the direct sum of these m copies.

(2) By Example 15.2.3, $\operatorname{Ind}_{A_n}^{S_n} \mathbb{1}_{A_n}$ is the permutation representation on S_n/A_n, which has two elements since $[S_n : A_n] = 2$. Its basis e_1, e_2 is indexed by the two cosets; even permutations fix both basis vectors, and odd permutations swap them. So $\operatorname{span}(e_1 + e_2)$ is a trivial subrepresentation, $\operatorname{span}(e_1 - e_2)$ is a copy of the sign representation, and the representation is $\mathbb{1} \oplus \mathbb{1}'$.

(3) By Frobenius reciprocity, the multiplicity of an irreducible X of S_5 in $\operatorname{Ind} V_4$ is $\langle \chi_{V_4}, \chi_{\operatorname{Res} X} \rangle_{S_4}$. The conjugacy classes of $S_4 \subset S_5$ are those of $e, (12), (12)(34), (123), (1234)$, of sizes $1, 6, 3, 8, 6$, and $\chi_{V_4} = (3, 1, -1, 0, -1)$ on these classes. Reading the character table of S_5 (Section 14.3) at these elements:

	e	(12)	$(12)(34)$	(123)	(1234)
$\mathbb{1}$	1	1	1	1	1
$\mathbb{1}'$	1	-1	1	1	-1
V	4	2	0	1	0
V'	4	-2	0	1	0
$\bigwedge^2 V$	6	0	-2	0	0
W	5	1	1	-1	-1
W'	5	-1	1	-1	1

Weighting by the class sizes,

$$\begin{aligned}
\langle \chi_{V_4}, \chi_{\operatorname{Res} \mathbb{1}} \rangle_{S_4} &= \tfrac{1}{24}(3 + 6 - 3 + 0 - 6) = 0, \\
\langle \chi_{V_4}, \chi_{\operatorname{Res} \mathbb{1}'} \rangle_{S_4} &= \tfrac{1}{24}(3 - 6 - 3 + 0 + 6) = 0, \\
\langle \chi_{V_4}, \chi_{\operatorname{Res} V} \rangle_{S_4} &= \tfrac{1}{24}(12 + 12 + 0 + 0 + 0) = 1, \\
\langle \chi_{V_4}, \chi_{\operatorname{Res} V'} \rangle_{S_4} &= \tfrac{1}{24}(12 - 12 + 0 + 0 + 0) = 0, \\
\langle \chi_{V_4}, \chi_{\operatorname{Res} \bigwedge^2 V} \rangle_{S_4} &= \tfrac{1}{24}(18 + 0 + 6 + 0 + 0) = 1, \\
\langle \chi_{V_4}, \chi_{\operatorname{Res} W} \rangle_{S_4} &= \tfrac{1}{24}(15 + 6 - 3 + 0 + 6) = 1, \\
\langle \chi_{V_4}, \chi_{\operatorname{Res} W'} \rangle_{S_4} &= \tfrac{1}{24}(15 - 6 - 3 + 0 - 6) = 0.
\end{aligned}$$

So $\operatorname{Ind}_{S_4}^{S_5} V_4 \cong V \oplus \bigwedge^2 V \oplus W$, of dimension $4 + 6 + 5 = 15 = [S_5 : S_4] \cdot 3$.

(4) By Proposition 15.2.6,

$$\langle \chi_{\operatorname{Ind} W}, \chi_V \rangle_G = \frac{1}{|G|\,|H|} \sum_{s \in G} \sum_{\substack{g \in G \\ s^{-1}gs \in H}} \overline{\chi_W(s^{-1}gs)}\, \chi_V(g).$$

For fixed s, the map $g \mapsto h = s^{-1}gs$ is a bijection from $\{g \in G : s^{-1}gs \in H\}$ to H, and $\chi_V(g) = \chi_V(shs^{-1}) = \chi_V(h)$ because χ_V is a class function on G.

So the inner sum equals $\sum_{h\in H} \overline{\chi_W(h)}\chi_V(h)$ for every s, and

$$\langle \chi_{\mathrm{Ind}\, W}, \chi_V\rangle_G = \frac{1}{|G|\,|H|}\cdot |G| \sum_{h\in H} \overline{\chi_W(h)}\,\chi_{\mathrm{Res}\, V}(h) = \langle \chi_W, \chi_{\mathrm{Res}\, V}\rangle_H .$$

(5) Let $\{\sigma_i\}$ be representatives of the cosets of H in G and $\{\tau_j\}$ representatives of the cosets of K in H. Every $g \in G$ can be written uniquely as $g = \sigma_i h$ with $h \in H$, and then $h = \tau_j k$ uniquely with $k \in K$; so $\{\sigma_i\tau_j\}$ is a set of representatives of the cosets of K in G. Let $X = \mathrm{Ind}_H^G(\mathrm{Ind}_K^H W)$. It contains $\mathrm{Ind}_K^H W = \bigoplus_j \tau_j W$, which contains the K-invariant subspace W, and

$$X = \bigoplus_i \sigma_i \Big(\bigoplus_j \tau_j W\Big) = \bigoplus_{i,j} (\sigma_i\tau_j) W.$$

So X is induced by W as a representation of G, and $X \cong \mathrm{Ind}_K^G W$ by Theorem 15.2.2.

CHAPTER 16

Real and Quaternionic Representations

So far we have studied actions of finite groups on complex vector spaces. In this chapter we consider actions on real vector spaces and how they sit among the complex representations. We will see that an irreducible complex representation falls into exactly one of three classes: it comes from a real representation, it carries a quaternionic structure, or its character is not real-valued. A number computed from the character, the Frobenius–Schur indicator, tells which.

Throughout, G is a finite group and all representations are finite-dimensional.

16.1 Representations over $\mathbb{R}$

Definition 16.1.1 (Representation over $\mathbb{R}$)

A ***representation of*** G ***over*** $\mathbb{R}$ is a finite-dimensional real vector space V_0 together with a group homomorphism $G \to \mathrm{GL}(V_0)$, where $\mathrm{GL}(V_0)$ is the group of invertible $\mathbb{R}$-linear maps $V_0 \to V_0$.

The averaging argument of Lemma 13.2.1 works over $\mathbb{R}$. If $\langle \cdot, \cdot \rangle_0$ is any inner product on V_0, then

$$\langle v, w \rangle = \frac{1}{|G|} \sum_{g \in G} \langle gv, gw \rangle_0$$

is a G-invariant inner product. If $W_0 \subset V_0$ is a G-invariant subspace, then so is its orthogonal complement $W_0^{\perp}$, exactly as in the proof of Maschke's theorem (Theorem 13.2.2). Hence every representation of a finite group over $\mathbb{R}$ is a direct sum of irreducible ones.

Part (i) of Schur's lemma (Lemma 13.3.1) also holds over $\mathbb{R}$, with the same proof: a nonzero equivariant map between irreducible representations is an isomorphism. Part (ii) fails, because its proof needs an eigenvalue.

Example 16.1.2 (Rotations of the plane)
Let $n \geqslant 3$, and let $\mathbb{Z}/n$ act on $\mathbb{R}^2$ by letting 1 act by the rotation

$$r = \begin{pmatrix} c & -s \\ s & c \end{pmatrix}, \qquad c = \cos\tfrac{2\pi}{n}, \quad s = \sin\tfrac{2\pi}{n}.$$

Since $s \neq 0$, the rotation r has no real eigenvalue, so no line is invariant and the representation is irreducible. A direct computation shows that the matrices commuting with r are those of the form $\left(\begin{smallmatrix} a & -b \\ b & a \end{smallmatrix}\right)$, so

$$\mathrm{End}_{\mathbb{Z}/n}(\mathbb{R}^2) = \left\{ \begin{pmatrix} a & -b \\ b & a \end{pmatrix} : a, b \in \mathbb{R} \right\} \cong \mathbb{C},$$

which is larger than $\mathbb{R} \cdot \mathrm{id}$.

16.2 Complexification

The main tool for relating real and complex representations is complexification (Section 7.10).

Definition 16.2.1 (Complexification of a representation)

Let V_0 be a representation of G over $\mathbb{R}$. Its ***complexification*** is the complex vector space $V = (V_0)_{\mathbb{C}} = V_0 \oplus iV_0$ (Definition 7.10.1), on which $g \in G$ acts by the complexified operator

$$g(v + iw) = gv + i\,gw \qquad (v, w \in V_0).$$

A real basis of V_0 is a complex basis of V, and in such a basis each g has the same real matrix on V as on V_0. Equivariant maps complexify as well, so complexification is a functor from representations of G over $\mathbb{R}$ to complex representations of G. The complexification is also written $V_0 \otimes_{\mathbb{R}} \mathbb{C}$.

Definition 16.2.2 (Real representation)

A complex representation V of G is ***real*** if $V \cong (V_0)_{\mathbb{C}}$ for some representation V_0 of G over $\mathbb{R}$.

Proposition 16.2.3 (Real representations have real characters)
If V is real, then $\chi_V(g) \in \mathbb{R}$ for every $g \in G$.

Proof. In a basis of V coming from a real basis of V_0, every g acts by a real matrix, whose trace is real. □

Complexification need not preserve irreducibility.

Example 16.2.4 (A real irreducible representation whose complexification is reducible)
The rotation representation of $\mathbb{Z}/n$ on $\mathbb{R}^2$ from Example 16.1.2 is irreducible over $\mathbb{R}$, but its complexification $\mathbb{C}^2$ is not. The vectors $(1,-i)$ and $(1,i)$ are eigenvectors of r with eigenvalues $e^{2\pi i/n}$ and $e^{-2\pi i/n}$, so

$$(\mathbb{R}^2)_{\mathbb{C}} \cong \chi \oplus \overline{\chi},$$

where χ is the 1-dimensional representation with $\chi(1) = e^{2\pi i/n}$. For $n \geqslant 3$ neither summand is real, since χ and $\overline{\chi}$ take nonreal values.

The converse of Proposition 16.2.3 is false, even for irreducible representations.

Example 16.2.5 (The quaternion group)
The quaternion group $Q_8 = \{\pm 1, \pm i, \pm j, \pm k\}$, with $i^2 = j^2 = k^2 = ijk = -1$, acts on $\mathbb{C}^2$ by

$$\pm 1 \mapsto \pm I, \quad \pm i \mapsto \pm \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \quad \pm j \mapsto \pm \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad \pm k \mapsto \pm \begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix}.$$

(These matrices multiply according to the rules of Q_8: each of the last three squares to $-I$, and the product of the matrices of i and j is the matrix of k.) The representation is faithful. Its character is $\chi(\pm 1) = \pm 2$ and $\chi = 0$ on the other six elements, so χ is real-valued, and it is irreducible, since $\langle \chi, \chi \rangle = \frac{1}{8}(4+4) = 1$.
This representation is not real. Suppose $V = \mathbb{C}^2$ were isomorphic to $(V_0)_{\mathbb{C}}$. Then in a suitable basis the elements of Q_8 would act by real matrices, so V_0 would be a faithful 2-dimensional representation over $\mathbb{R}$. Averaging gives a Q_8-invariant inner product on V_0, and in an orthonormal basis we get an injective homomorphism $Q_8 \to O(2)$. The image of $-1 = i^2$ is the square of an element of $O(2)$, so it has determinant 1, and it has order 2; so it is the rotation by π, which is $-I$. Then the images of $\pm i, \pm j, \pm k$ are six distinct

elements of $O(2)$ whose square is $-I$. But a reflection squares to I, and a rotation by θ squares to $-I$ only for $\theta = \pm\pi/2$; so $O(2)$ contains only two such elements, a contradiction.

We will see in Section 16.5 that this representation is quaternionic, and that an irreducible representation with real character is always either real or quaternionic.

16.3 Characterizing Real Representations

A map $\tau : V \to V$ of complex vector spaces is ***antilinear*** (or conjugate-linear) if $\tau(v+w) = \tau(v) + \tau(w)$ and $\tau(\lambda v) = \bar{\lambda}\,\tau(v)$ for all $v, w \in V$ and $\lambda \in \mathbb{C}$. The composite of two antilinear maps is linear, and the composite of a linear and an antilinear map is antilinear.

Proposition 16.3.1 (Antilinear involution criterion)

A complex representation V of G is real if and only if there is a G-equivariant antilinear map $\tau : V \to V$ with $\tau^2 = \mathrm{id}$.

Proof. ($\Rightarrow$) If $V = (V_0)_{\mathbb{C}}$, let $\tau(v + iw) = v - iw$ for $v, w \in V_0$: complex conjugation with respect to V_0. Then τ is antilinear, $\tau^2 = \mathrm{id}$, and τ is G-equivariant because each g preserves V_0.

($\Leftarrow$) Let $V_0 = \{v \in V : \tau(v) = v\}$ and $V_1 = \{v \in V : \tau(v) = -v\}$. These are $\mathbb{R}$-subspaces of V (not $\mathbb{C}$-subspaces), and every $v \in V$ decomposes as

$$v = \underbrace{\frac{v + \tau(v)}{2}}_{\in V_0} + \underbrace{\frac{v - \tau(v)}{2}}_{\in V_1},$$

with $V_0 \cap V_1 = 0$. Since $\tau(iv) = -i\tau(v)$, multiplication by i maps V_0 to V_1, and multiplication by $-i$ maps V_1 back to V_0; so $V_1 = iV_0$ and $V = V_0 \oplus iV_0$. Hence the map $(V_0)_{\mathbb{C}} \to V$, $v + iw \mapsto v + iw$, is a $\mathbb{C}$-linear isomorphism. Since τ commutes with every g, each g preserves V_0, which is therefore a representation of G over $\mathbb{R}$, and the isomorphism $(V_0)_{\mathbb{C}} \to V$ is G-equivariant. □

16.4 Invariant Bilinear Forms

A bilinear form $B : V \times V \to \mathbb{C}$ on a representation V is G-***invariant*** if $B(gv, gw) = B(v, w)$ for all $g \in G$ and $v, w \in V$. Recall from Chapter 9 that a linear map $\varphi : V \to V^*$ determines the bilinear form $B(v, w) = \varphi(v)(w)$, that every bilinear form arises in this way from exactly one φ, and that B is nondegenerate if and only if φ is an isomorphism.

Lemma 16.4.1 (Invariant forms and equivariant maps)

Let V^* carry the dual representation, $(g\ell)(w) = \ell(g^{-1}w)$. The bilinear form B determined by $\varphi : V \to V^*$ is G-invariant if and only if φ is G-equivariant. So the G-invariant bilinear forms on V correspond to the elements of $\mathrm{Hom}_G(V, V^*)$.

Proof. For $g \in G$ and $v, w \in V$, we have $\varphi(gv)(w) = B(gv, w)$ and $(g\varphi(v))(w) = \varphi(v)(g^{-1}w) = B(v, g^{-1}w)$. So φ is equivariant if and only if $B(gv, w) = B(v, g^{-1}w)$ for all g, v, w; substituting $w = gw'$, this says $B(gv, gw') = B(v, w')$ for all g, v, w'. □

Proposition 16.4.2 (Invariant forms on an irreducible representation)

Let V be an irreducible representation of G.

(i) $V \cong V^*$ if and only if χ_V is real-valued.

(ii) If χ_V is not real-valued, the only G-invariant bilinear form on V is 0.

(iii) If χ_V is real-valued, V has a nonzero G-invariant bilinear form B, unique up to a scalar factor. Every nonzero G-invariant bilinear form on V is nondegenerate, and B is either symmetric ($B(w, v) = B(v, w)$) or skew-symmetric ($B(w, v) = -B(v, w)$).

Proof. (i) We have $\chi_{V^*} = \overline{\chi_V}$ (Proposition 13.4.3), and two representations are isomorphic if and only if they have the same character (Corollary 13.5.6). So $V \cong V^*$ if and only if $\chi_V = \overline{\chi_V}$.

(ii) and (iii) The dual V^* is irreducible, since $\langle \chi_{V^*}, \chi_{V^*} \rangle = \langle \overline{\chi_V}, \overline{\chi_V} \rangle = \langle \chi_V, \chi_V \rangle = 1$. By Lemma 16.4.1, invariant forms correspond to elements of $\mathrm{Hom}_G(V, V^*)$. By Schur's lemma this space is 0 if $V \not\cong V^*$. If $\psi : V \to V^*$ is an isomorphism of representations, then every $\varphi \in \mathrm{Hom}_G(V, V^*)$ satisfies $\psi^{-1}\varphi \in \mathrm{Hom}_G(V, V) = \mathbb{C} \cdot \mathrm{id}$, so $\mathrm{Hom}_G(V, V^*) = \mathbb{C}\psi$ is 1-dimensional. A nonzero φ is an isomorphism by Schur's lemma, so every nonzero invariant form is nondegenerate. (Directly: $\{v : B(v, w) = 0 \text{ for all } w\}$ is a subrepresentation, since $B(gv, w) = B(v, g^{-1}w)$, and it is not all of V when $B \neq 0$, so it is 0.)

Finally, write $B = B^+ + B^-$ with

$$B^{\pm}(v, w) = \tfrac{1}{2}\big(B(v, w) \pm B(w, v)\big).$$

Both B^+ (symmetric) and B^- (skew-symmetric) are G-invariant bilinear forms, so $B^+ = aB$ and $B^- = bB$ for some $a, b \in \mathbb{C}$ with $a + b = 1$. If a and b were both nonzero, then $B = a^{-1}B^+$ would be symmetric and $B = b^{-1}B^-$ skew-symmetric, so $B(v, w) = B(w, v) = -B(v, w)$ and $B = 0$, a contradiction. Hence $B = B^+$ or $B = B^-$. □

Proposition 16.4.3 (Real if and only if symmetric)
An irreducible complex representation V of G is real if and only if V carries a nondegenerate G-invariant symmetric bilinear form $B : V \times V \to \mathbb{C}$.

Proof. ($\Rightarrow$) Let $V = (V_0)_{\mathbb{C}}$. By averaging, V_0 has a G-invariant inner product B_0. Extend it $\mathbb{C}$-bilinearly:

$$B(v_1 + iw_1, v_2 + iw_2) = B_0(v_1, v_2) - B_0(w_1, w_2) + i\big(B_0(w_1, v_2) + B_0(v_1, w_2)\big).$$

This is a symmetric, G-invariant, $\mathbb{C}$-bilinear form on V. It is nondegenerate: if $B(v + iw, u) = 0$ for all $u \in V_0$, then $B_0(v, u) + iB_0(w, u) = 0$, so $B_0(v, u) = B_0(w, u) = 0$ for all $u \in V_0$, and $v = w = 0$.

($\Leftarrow$) Choose a G-invariant Hermitian inner product $\langle \cdot, \cdot \rangle$ on V (Lemma 13.2.1), conjugate-linear in the first variable. The map $v \mapsto \langle v, \cdot \rangle$ from V to V^* is antilinear and injective, hence bijective, and it is G-equivariant: $\langle gv, w \rangle = \langle v, g^{-1}w \rangle$. The form B gives a G-equivariant linear isomorphism $\varphi : V \to V^*$, $\varphi(v) = B(v, \cdot)$. Composing φ with the inverse of $v \mapsto \langle v, \cdot \rangle$ gives a G-equivariant antilinear bijection $\tau : V \to V$, characterized by

$$\langle \tau(v), w \rangle = B(v, w) \qquad \text{for all } v, w \in V.$$

The map τ^2 is $\mathbb{C}$-linear and G-equivariant, so by Schur's lemma $\tau^2 = \lambda \cdot \mathrm{id}$ for some $\lambda \in \mathbb{C}$. For $v \neq 0$,

$$\overline{\lambda}\, \langle v, v \rangle = \langle \tau^2(v), v \rangle = B(\tau(v), v) = B(v, \tau(v)) = \langle \tau(v), \tau(v) \rangle > 0,$$

using the symmetry of B in the third equality and $\tau(v) \neq 0$ (as τ is bijective) at the end. So λ is real and positive. The map $\lambda^{-1/2}\tau$ is still antilinear and G-equivariant, and $(\lambda^{-1/2}\tau)^2 = \lambda^{-1}\tau^2 = \mathrm{id}$ because $\lambda^{-1/2}$ is real. By Proposition 16.3.1, V is real. □

16.5 Quaternionic Representations

Definition 16.5.1 (Quaternionic structure)

A ***quaternionic structure*** on a complex vector space V is an antilinear map $J : V \to V$ with $J^2 = -\mathrm{id}$. A complex representation V of G is ***quaternionic*** if it has a G-equivariant quaternionic structure.

The name comes from the quaternions

$$\mathbb{H} = \{a + bi + cj + dk : a, b, c, d \in \mathbb{R}\}, \qquad i^2 = j^2 = k^2 = ijk = -1,$$

a ***division ring***: a ring in which $1 \neq 0$ and every nonzero element has a multiplicative inverse. Unlike a field, $\mathbb{H}$ is not commutative. As a real vector space $\mathbb{H} = \mathbb{C} \oplus j\mathbb{C}$, where $\mathbb{C} = \mathbb{R} \oplus \mathbb{R}i$, and $j\lambda = \overline{\lambda}j$ for $\lambda \in \mathbb{C}$. A quaternionic structure J makes V into a module over $\mathbb{H}$ (Definition 10.1.1), with i acting by multiplication by i and j acting by J: the relations $j^2 = -1$ and $ji = -ij$ hold because $J^2 = -\mathrm{id}$ and $J(iv) = -iJ(v)$. Conversely, an $\mathbb{H}$-module is a complex vector space together with such a J.

Proposition 16.5.2 (Skew-symmetric forms give quaternionic structures)
If an irreducible representation V of G carries a nonzero G-invariant skew-symmetric bilinear form B, then V is quaternionic.

Proof. By Proposition 16.4.2, B is nondegenerate. Exactly as in the proof of Proposition 16.4.3, there is a G-equivariant antilinear bijection $J : V \to V$ with $\langle J(v), w\rangle = B(v, w)$, and $J^2 = \lambda \cdot \mathrm{id}$ for some $\lambda \in \mathbb{C}$. Now skew-symmetry gives, for $v \neq 0$,

$$\overline{\lambda}\,\langle v, v\rangle = \langle J^2(v), v\rangle = B(J(v), v) = -B(v, J(v)) = -\langle J(v), J(v)\rangle < 0,$$

so λ is real and negative. Replacing J by $|\lambda|^{-1/2}J$ gives $J^2 = -\mathrm{id}$. □

Proposition 16.5.3 (Real and quaternionic representations)

(i) The character of a real or quaternionic representation is real-valued.
(ii) An irreducible representation cannot be both real and quaternionic.

Proof. (i) Let $K : V \to V$ be a G-equivariant antilinear bijection (τ or J). If $gv = \lambda v$, then $g(Kv) = K(gv) = K(\lambda v) = \overline{\lambda}\,Kv$. So K maps the λ-eigenspace E_λ of g injectively into $E_{\overline{\lambda}}$, and likewise $E_{\overline{\lambda}}$ into E_λ; as K is $\mathbb{R}$-linear, E_λ and $E_{\overline{\lambda}}$ have the same real dimension, hence the same complex dimension. Since g has finite order it is diagonalizable (Lemma 13.1.5), so $\chi_V(g) = \sum_\lambda \lambda \dim E_\lambda$, which equals its own complex conjugate.

(ii) Suppose V is irreducible and has a G-equivariant antilinear τ with $\tau^2 = \mathrm{id}$ and a G-equivariant antilinear J with $J^2 = -\mathrm{id}$. Then τJ is $\mathbb{C}$-linear and G-equivariant, so $\tau J = c \cdot \mathrm{id}$ for some $c \in \mathbb{C}$ by Schur's lemma, and $J = \tau(\tau J) = \tau \circ (c \cdot \mathrm{id}) = \overline{c}\,\tau$. But then $J^2(v) = \overline{c}\,\tau(\overline{c}\,\tau(v)) = c\,\overline{c}\,\tau^2(v) = |c|^2 v$, contradicting $J^2 = -\mathrm{id}$. □

Theorem 16.5.4 (Real, quaternionic, or neither)

Let V be an irreducible complex representation of G. Exactly one of the following holds.

(i) χ_V is not real-valued. Then $V \not\cong V^*$, V is neither real nor quaternionic, and V has no nonzero G-invariant bilinear form.

(ii) V is real. Then the nonzero G-invariant bilinear forms on V are symmetric and nondegenerate.

(iii) V is quaternionic. Then the nonzero G-invariant bilinear forms on V are skew-symmetric and nondegenerate.

Proof. If χ_V is not real-valued, Propositions 16.4.2 and 16.5.3(i) give (i). If χ_V is real-valued, then by Proposition 16.4.2 the G-invariant bilinear forms on V are the multiples of one nonzero form B, which is nondegenerate and either symmetric or skew-symmetric. If B is symmetric, V is real by Proposition 16.4.3; if B is skew-symmetric, V is quaternionic by Proposition 16.5.2. By Proposition 16.5.3(ii) these cases exclude each other, so V is real exactly when B is symmetric and quaternionic exactly when B is skew-symmetric. □

Proposition 16.5.5 (Quaternionic representations have even dimension)

If V is a quaternionic representation, then $\dim_{\mathbb{C}} V$ is even. Consequently, an irreducible representation with real-valued character and odd dimension is real.

Proof. Let $n = \dim_{\mathbb{C}} V$ and choose a basis $e_1, \dots, e_n$ of V. Let M be the matrix whose kth column is the coordinate vector of $J(e_k)$. By antilinearity, $J(\sum_k z_k e_k) = \sum_k \overline{z_k}\, J(e_k)$, so in coordinates $J(z) = M\bar{z}$, and $J^2(z) = M\,\overline{M\bar{z}} = M\overline{M}z$. Hence $M\overline{M} = -I$, and taking determinants, $|\det M|^2 = \det M \cdot \overline{\det M} = (-1)^n$. The left side is positive, so n is even. The second statement follows from Theorem 16.5.4. □

For an irreducible quaternionic representation there is also a second proof: by Theorem 16.5.4, V carries a nondegenerate skew-symmetric bilinear form, and such forms exist only in even dimension (Section 9.7).

16.6 The Frobenius–Schur Indicator

The trichotomy of Theorem 16.5.4 can be detected numerically, from the character alone.

Definition 16.6.1 (Frobenius–Schur indicator)

The ***Frobenius–Schur indicator*** of a representation V of G is

$$\nu(V) = \frac{1}{|G|} \sum_{g \in G} \chi_V(g^2).$$

Theorem 16.6.2 (Frobenius–Schur)

For an irreducible representation V of G,

$$\nu(V) = \begin{cases} 1 & \text{if } V \text{ is real,} \\ 0 & \text{if } \chi_V \text{ is not real-valued,} \\ -1 & \text{if } V \text{ is quaternionic.} \end{cases}$$

Proof. Let $s : V \otimes V \to V \otimes V$ be the linear map with $s(v \otimes w) = w \otimes v$. It commutes with the action of G and satisfies $s^2 = \mathrm{id}$, so $V \otimes V = S \oplus A$, where

$$S = \{x : s(x) = x\} \quad \text{and} \quad A = \{x : s(x) = -x\}$$

are subrepresentations. The quotient maps $V \otimes V \to \mathrm{Sym}^2 V$ and $V \otimes V \to \bigwedge^2 V$ of Definitions 9.10.3 and 9.10.4 restrict to isomorphisms $S \cong \mathrm{Sym}^2 V$ and $A \cong \bigwedge^2 V$: the first map kills exactly the span of the tensors $x - s(x)$, which is A, and the second kills exactly the span of the tensors $v \otimes v$, which is S, because $v \otimes w + w \otimes v = (v + w) \otimes (v + w) - v \otimes v - w \otimes w$.

Step 1: $\nu(V) = \dim S^G - \dim A^G$. Fix $g \in G$ and a basis $e_1, \dots, e_n$ of V consisting of eigenvectors of g (Lemma 13.1.5), with eigenvalues $\lambda_1, \dots, \lambda_n$. Then S has the basis $e_i \otimes e_j + e_j \otimes e_i$ $(i \leqslant j)$ and A has the basis $e_i \otimes e_j - e_j \otimes e_i$ $(i < j)$, and each of these vectors is an eigenvector of g with eigenvalue $\lambda_i \lambda_j$. Hence

$$\chi_S(g) - \chi_A(g) = \sum_{i \leqslant j} \lambda_i \lambda_j - \sum_{i < j} \lambda_i \lambda_j = \sum_i \lambda_i^2 = \chi_V(g^2).$$

Averaging over G and using $\dim X^G = \frac{1}{|G|} \sum_g \chi_X(g)$ (Proposition 13.5.1) gives Step 1.

Step 2: invariant forms. A bilinear form B on V is the same as a linear functional β on $V \otimes V$, via $\beta(v \otimes w) = B(v, w)$ (Theorem 9.8.6). The form B is G-invariant if and only if $\beta(gx) = \beta(x)$ for all g and x, that is, if and only if β is a G-invariant element of the dual representation $(V \otimes V)^*$. The form B is symmetric if and only if $\beta \circ s = \beta$, that is, β vanishes on A; and B is skew-symmetric if and only if β vanishes on S. Since $V \otimes V = S \oplus A$, restriction to S identifies the symmetric

forms with S^*, and restriction to A identifies the skew-symmetric forms with A^*, compatibly with the actions of G. Since $\chi_{S^*} = \overline{\chi_S}$,

$$\dim(S^*)^G = \frac{1}{|G|}\sum_{g\in G}\overline{\chi_S(g)} = \overline{\dim S^G} = \dim S^G,$$

and likewise $\dim(A^*)^G = \dim A^G$. So $\nu(V) = s_+ - s_-$, where s_+ is the dimension of the space of G-invariant symmetric bilinear forms on V and s_- that of the space of G-invariant skew-symmetric ones.

Step 3. By Theorem 16.5.4: if χ_V is not real-valued, $s_+ = s_- = 0$ and $\nu(V) = 0$. If V is real, the invariant forms are the multiples of one symmetric form, so $s_+ = 1$, $s_- = 0$, and $\nu(V) = 1$. If V is quaternionic, they are the multiples of one skew-symmetric form, so $s_+ = 0$, $s_- = 1$, and $\nu(V) = -1$. □

To compute $\nu(V)$ from a character table, one needs to know, for each conjugacy class, the class of the squares of its elements.

16.7 Examples

Example 16.7.1 (The standard representation of S_3)

The 2-dimensional irreducible representation V of S_3 is real. This can be seen in three ways.

- Directly: $S_3 \cong D_3$ acts on $V_0 = \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_1+x_2+x_3 = 0\} \cong \mathbb{R}^2$ by permuting coordinates (as the rotations and reflections of an equilateral triangle), and $V = (V_0)_{\mathbb{C}}$.
- Through invariant forms: in the notation of the proof of Theorem 16.6.2, $S = \mathrm{Sym}^2 V \cong U \oplus V$ contains the trivial representation once, while $A = \bigwedge^2 V \cong U'$ does not contain it (Example 13.3.6). So V has an invariant symmetric form and no invariant skew-symmetric form.
- Through the indicator: squaring sends e and the three transpositions to e, and the two 3-cycles to 3-cycles, so

$$\nu(V) = \tfrac{1}{6}\big(4\cdot\chi_V(e) + 2\cdot\chi_V((123))\big) = \tfrac{1}{6}(8-2) = 1.$$

Example 16.7.2 (The quaternion group, again)

The 2-dimensional representation of Q_8 in Example 16.2.5 is quaternionic. Identify $\mathbb{C}^2$ with $\mathbb{H}$ by $(z_1, z_2) \leftrightarrow z_1 + jz_2$, where $\mathbb{C} = \mathbb{R}\oplus\mathbb{R}i \subset \mathbb{H}$, and let $\mathbb{C}$ act on $\mathbb{H}$ by multiplication on the right: $(z_1 + jz_2)\lambda = z_1\lambda + j(z_2\lambda)$, which is the usual scalar multiplication on $\mathbb{C}^2$. Left multiplication by an element of

Q_8 commutes with right multiplication, so it is $\mathbb{C}$-linear, and its matrix is the one in Example 16.2.5. For instance, using $ij = -ji$, $ji = -k$, and $kj = -i$,

$$\begin{aligned} i(z_1 + jz_2) &= iz_1 + j(-iz_2), \\ j(z_1 + jz_2) &= -z_2 + jz_1, \\ k(z_1 + jz_2) &= -iz_2 + j(-iz_1). \end{aligned}$$

Right multiplication by j,

$$J(z_1 + jz_2) = (z_1 + jz_2)j = -\overline{z_2} + j\overline{z_1}, \quad \text{that is,} \quad J(z_1, z_2) = (-\overline{z_2}, \overline{z_1}),$$

is antilinear (because $\lambda j = j\overline{\lambda}$ for $\lambda \in \mathbb{C}$), satisfies $J^2 = -\mathrm{id}$, and commutes with left multiplication by associativity. So J is a Q_8-equivariant quaternionic structure. Consistently with Theorem 16.6.2, $\nu = -1$ (Exercise 16.7.3(4)).

Exercise 16.7.3 (Exercises)

(1) Determine which irreducible representations of S_4 are real, which are quaternionic, and which are neither.

(2) For $h \in G$, let $r(h)$ be the number of $g \in G$ with $g^2 = h$. Show that r is a class function and that $r = \sum_V \nu(V)\chi_V$, the sum running over the irreducible representations of G. Deduce that $\sum_V \nu(V) \dim V$ is the number of elements $g \in G$ with $g^2 = e$. Check this for S_3 and for Q_8.

(3) Let $|G|$ be odd. Show that the trivial representation is the only irreducible representation of G whose character is real-valued. (Hint: $g \mapsto g^2$ is a bijection of G.)

(4) Compute the Frobenius–Schur indicator of the 2-dimensional irreducible representation of Q_8.

(5) Show that the 3-dimensional irreducible representations Y and Z of A_5 are real.

Solution. **(1)** All characters of S_4 are real-valued (see its character table in Section 14.1), so each irreducible representation is real or quaternionic. The irreducible representations U, U', V, V' have odd dimension $1, 1, 3, 3$, so they are real by Proposition 16.5.5. For W we compute the indicator. On the classes of $e, (12), (12)(34), (123), (1234)$, of sizes $1, 6, 3, 8, 6$, we have $\chi_W = (2, 0, 2, -1, 0)$. Squaring sends $e, (12), (12)(34)$ to e, a 3-cycle to a 3-cycle, and a 4-cycle to a double transposition, so $\chi_W(g^2) = (2, 2, 2, -1, 2)$

and

$$\nu(W) = \tfrac{1}{24}(2 + 12 + 6 - 8 + 12) = 1.$$

So W is real as well. (Alternatively, W is the standard representation of S_3 pulled back along $S_4 \to S_4/K \cong S_3$, and that representation is real.) Every irreducible representation of S_4 is real.

(2) If $g^2 = h$, then $(xgx^{-1})^2 = xhx^{-1}$, so $g \mapsto xgx^{-1}$ is a bijection from the square roots of h to those of xhx^{-1}, and r is a class function. Since $\sum_h \overline{\chi_V(h)}\, r(h) = \sum_{g\in G} \overline{\chi_V(g^2)}$,

$$\langle \chi_V, r\rangle = \frac{1}{|G|}\sum_{g\in G} \overline{\chi_V(g^2)} = \overline{\nu(V)} = \nu(V),$$

as $\nu(V)$ is an integer. The irreducible characters form an orthonormal basis of the class functions (Theorem 13.6.2), so $r = \sum_V \langle \chi_V, r\rangle \chi_V = \sum_V \nu(V)\chi_V$. Evaluating at e gives $\#\{g : g^2 = e\} = \sum_V \nu(V)\dim V$.

For S_3: U, U', V are real, so the sum is $1 + 1 + 2 = 4$, and indeed e and the three transpositions are the elements with $g^2 = e$. For Q_8: there are four 1-dimensional representations, pulled back from $Q_8/\{\pm 1\} \cong \mathbb{Z}/2 \times \mathbb{Z}/2$; they take the values ± 1 and are real. With the quaternionic 2-dimensional representation, the sum is $4 - 2 = 2$, and indeed only ± 1 square to 1.

(3) Let $|G| = m$ be odd. Since $g^m = e$, we have $(g^{(m+1)/2})^2 = g^{m+1} = g$, so squaring is surjective, hence bijective, on the finite set G. Therefore, for an irreducible V,

$$\nu(V) = \frac{1}{|G|}\sum_{g\in G} \chi_V(g^2) = \frac{1}{|G|}\sum_{g\in G} \chi_V(g) = \dim V^G.$$

If V is not trivial, then V^G is a proper subrepresentation of the irreducible V, so $V^G = 0$ and $\nu(V) = 0$. By Theorem 16.6.2, χ_V is not real-valued.

(4) Q_8 has five conjugacy classes, $\{1\}, \{-1\}, \{\pm i\}, \{\pm j\}, \{\pm k\}$, of sizes $1, 1, 2, 2, 2$, and $\chi = (2, -2, 0, 0, 0)$. Squaring sends ± 1 to 1 and the other six elements to -1, so $\chi(g^2) = (2, 2, -2, -2, -2)$ and

$$\nu = \tfrac{1}{8}\big(2 + 2 + 2\cdot(-2)\cdot 3\big) = \tfrac{1}{8}(4 - 12) = -1.$$

So the representation is quaternionic, as found in the text.

(5) The characters of Y and Z are real-valued (see the character table of A_5 in Section 14.4) and their dimension, 3, is odd, so they are real by Proposition 16.5.5. One of them is also directly the complexification of the real representation $A_5 \hookrightarrow SO(3)$ given by the rotational symmetries of the icosahedron.

PART V

Further Topics

CHAPTER 17

Further Topics in Algebra

This chapter collects three subjects that extend the algebra of the earlier chapters: the arithmetic of commutative rings, leading to unique factorization; Galois theory, which connects field extensions to group theory; and an introduction to Lie groups and Lie algebras, the continuous analogues of finite groups and their representations. These topics go beyond the course on which this book is based, and the chapter is a survey. Some of the deeper theorems are stated without proof, and we say so each time.

17.1 Commutative Rings

In this section all rings are commutative with identity (Definition 4.1.1).

17.1.1 Ideals and Quotient Rings

Recall from Chapter 10 that an ***ideal*** of a ring R is an additive subgroup $I \subseteq R$ such that $ra \in I$ for all $r \in R$ and $a \in I$. For $a \in R$, the ***principal ideal*** generated by a is $(a) = aR = \{ra : r \in R\}$. More generally, for $a_1, \dots, a_n \in R$, the set

$$(a_1, \dots, a_n) = \{r_1 a_1 + \cdots + r_n a_n : r_1, \dots, r_n \in R\}$$

is the smallest ideal containing $a_1, \dots, a_n$, the ideal ***generated*** by them. We say that a ***divides*** b, written $a \mid b$, if $b = ac$ for some $c \in R$; equivalently, $(b) \subseteq (a)$.

Definition 17.1.1 (Quotient ring)

Let I be an ideal of R. The ***quotient ring*** R/I is the quotient group of $(R, +)$ by I, whose elements are the cosets $a + I$, with the multiplication $(a + I)(b + I) = ab + I$.

The multiplication is well defined: if $a' = a + x$ and $b' = b + y$ with $x, y \in I$, then $a'b' = ab + (ay + xb + xy) \in ab + I$. The ring axioms for R/I follow

from those for R, and the projection $R \to R/I$, $a \mapsto a + I$, is a surjective ring homomorphism with kernel I.

Proposition 17.1.2 (Kernels and quotients)

Let $\varphi : R \to S$ be a ring homomorphism. Then $\ker \varphi$ is an ideal of R, $\operatorname{Im} \varphi$ is a subring of S, and $a + \ker \varphi \mapsto \varphi(a)$ is an isomorphism of rings $R/\ker \varphi \to \operatorname{Im} \varphi$.

Proof. The kernel is an additive subgroup, and if $a \in \ker \varphi$ and $r \in R$, then $\varphi(ra) = \varphi(r) \cdot 0 = 0$. The image is closed under addition and multiplication and contains $\varphi(1) = 1$. By the First Isomorphism Theorem (Theorem 3.3.3) for the additive groups, the map $a + \ker \varphi \mapsto \varphi(a)$ is a well-defined isomorphism of additive groups onto $\operatorname{Im} \varphi$. It respects multiplication because $\varphi(ab) = \varphi(a)\varphi(b)$, and it sends $1 + \ker \varphi$ to 1. □

Definition 17.1.3 (Integral domain)

A ring R with $1 \neq 0$ is an ***integral domain*** if it has no zero divisors: $ab = 0$ implies $a = 0$ or $b = 0$.

In an integral domain we may cancel nonzero factors: if $ab = ac$ and $a \neq 0$, then $a(b - c) = 0$, so $b = c$.

Definition 17.1.4 (Prime and maximal ideals)

Let $I \neq R$ be an ideal of R. Then I is ***prime*** if $ab \in I$ implies $a \in I$ or $b \in I$, and ***maximal*** if the only ideals J with $I \subseteq J \subseteq R$ are I and R.

Proposition 17.1.5 (Quotients by prime and maximal ideals)

Let $I \neq R$ be an ideal of R.

(i) R/I is an integral domain if and only if I is prime.

(ii) R/I is a field if and only if I is maximal.

In particular, every maximal ideal is prime.

Proof. Note that $1 + I \neq 0 + I$, since $1 \notin I$.

(i) In R/I, we have $(a + I)(b + I) = 0$ if and only if $ab \in I$, and $a + I = 0$ if and only if $a \in I$.

(ii) Suppose R/I is a field, and let $J \supsetneq I$ be an ideal. Choose $a \in J \setminus I$. Then $a + I$ has an inverse $b + I$, so $ab - 1 \in I \subseteq J$, and $1 = ab - (ab - 1) \in J$; hence

$J = R$. Conversely, suppose I is maximal and $a \notin I$. The ideal $I + (a) = \{x + ra : x \in I, r \in R\}$ strictly contains I, so it is R, and $1 = x + ab$ for some $x \in I$ and $b \in R$. Then $(a + I)(b + I) = 1 + I$.

The last statement holds because a field has no zero divisors. □

17.1.2 Divisibility in Integral Domains

Example 17.1.6 (Integral domains)

Every field is an integral domain, and so is every subring of a field. For instance $\mathbb{Z}$, the Gaussian integers $\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\}$, and $\mathbb{Z}[\sqrt{-5}] = \{a + b\sqrt{-5} : a, b \in \mathbb{Z}\}$ are subrings of $\mathbb{C}$, hence integral domains. If k is a field, then $k[x]$ is an integral domain, since $\deg(fg) = \deg f + \deg g$ for nonzero f, g. The ring $\mathbb{Z}/6$ is not an integral domain, since $2 \cdot 3 = 0$.

Definition 17.1.7 (Units, associates, irreducible and prime elements)

Let R be an integral domain.

- A ***unit*** is an element $u \in R$ with a multiplicative inverse: $uv = 1$ for some $v \in R$. The units form a group $R^\times$ under multiplication.
- Two elements a, b are ***associates*** if $a = ub$ for some unit u.
- A nonzero non-unit $p \in R$ is ***irreducible*** if $p = ab$ implies that a or b is a unit.
- A nonzero non-unit $p \in R$ is ***prime*** if $p \mid ab$ implies $p \mid a$ or $p \mid b$.

In an integral domain, a and b are associates if and only if $(a) = (b)$: if $a = bc$ and $b = ad$, then $a = acd$, so either $a = b = 0$ or $cd = 1$. A nonzero element p is prime if and only if (p) is a prime ideal.

Proposition 17.1.8 (Prime implies irreducible)

In an integral domain, every prime element is irreducible.

Proof. Suppose p is prime and $p = ab$. Then $p \mid ab$, so $p \mid a$ or $p \mid b$; say $a = pc$. Then $p = pcb$, and canceling $p \neq 0$ gives $1 = cb$, so b is a unit. □

NOTE The converse fails in general: in $\mathbb{Z}[\sqrt{-5}]$ the element 2 is irreducible but not prime (Example 17.1.18). It holds in principal ideal domains (Theorem 17.1.13).

17.1.3 Euclidean Domains

Definition 17.1.9 (Euclidean domain)

An integral domain R is a ***Euclidean domain*** if there is a function $N : R\setminus\{0\} \to \mathbb{Z}_{\geqslant 0}$, a ***Euclidean function***, such that for all $a, b \in R$ with $b \neq 0$ there exist $q, r \in R$ with

$$a = bq + r, \qquad \text{where } r = 0 \text{ or } N(r) < N(b).$$

Example 17.1.10 (Euclidean domains)

- $\mathbb{Z}$, with $N(a) = |a|$ (division with remainder).
- $k[x]$ for a field k, with $N(f) = \deg f$. This is long division: if a has leading term αx^m and b has leading term βx^n with $m \geqslant n$, then $a - \alpha\beta^{-1}x^{m-n}b$ has smaller degree than a, and repeating this gives $a = bq + r$ with $r = 0$ or $\deg r < \deg b$.
- The Gaussian integers $\mathbb{Z}[i]$, with $N(a+bi) = a^2 + b^2$ (Exercise 17.1.27(2)).
- The Eisenstein integers $\mathbb{Z}[\omega] = \{a + b\omega : a, b \in \mathbb{Z}\}$, where $\omega = e^{2\pi i/3}$, with $N(a + b\omega) = |a + b\omega|^2 = a^2 - ab + b^2$. The argument is the same as for $\mathbb{Z}[i]$, since every complex number lies within distance $1/\sqrt{3} < 1$ of an element of $\mathbb{Z}[\omega]$.

17.1.4 Principal Ideal Domains

Definition 17.1.11 (Principal ideal domain)

An integral domain R is a ***principal ideal domain*** (PID) if every ideal of R is principal, that is, of the form (a) for some $a \in R$.

Theorem 17.1.12 (Every Euclidean domain is a PID)

If R is a Euclidean domain, then R is a principal ideal domain.

Proof. Let N be a Euclidean function and $I \subseteq R$ an ideal. If $I = 0$, then $I = (0)$. Otherwise choose $b \in I \setminus \{0\}$ with $N(b)$ minimal. Clearly $(b) \subseteq I$. For $a \in I$, write $a = bq + r$ with $r = 0$ or $N(r) < N(b)$. Since $r = a - bq \in I$, minimality forces $r = 0$, so $a = bq \in (b)$. □

In particular $\mathbb{Z}$, $k[x]$, $\mathbb{Z}[i]$ and $\mathbb{Z}[\omega]$ are principal ideal domains.

NOTE The converse fails: the ring $\mathbb{Z}\left[\frac{1+\sqrt{-19}}{2}\right]$ is a principal ideal domain that has no Euclidean function (Motzkin, 1949).

Theorem 17.1.13 (In a PID, irreducible $\iff$ prime)

Let R be a principal ideal domain and $p \in R$ a nonzero non-unit. Then p is irreducible if and only if p is prime.

Proof. ($\Leftarrow$) This is Proposition 17.1.8.

($\Rightarrow$) Suppose p is irreducible and $p \mid ab$. Since R is a PID, $(p, a) = (d)$ for some d. Then $d \mid p$, so $p = de$, and since p is irreducible, either d is a unit or e is a unit. If e is a unit, then $(d) = (p)$, so $a \in (p)$ and $p \mid a$. If d is a unit, then $(p, a) = R$, so $1 = rp + sa$ for some $r, s \in R$, and $b = rpb + sab$; both terms are divisible by p, so $p \mid b$. □

Corollary 17.1.14 (Irreducible elements generate maximal ideals)

If R is a principal ideal domain and $p \in R$ is irreducible, then (p) is a maximal ideal, so $R/(p)$ is a field.

Proof. Since p is not a unit, $(p) \neq R$. If $(p) \subseteq J \subseteq R$ for an ideal $J = (d)$, then $d \mid p$, so d is a unit (and $J = R$) or an associate of p (and $J = (p)$). The last statement is Proposition 17.1.5. □

17.1.5 Unique Factorization Domains

Definition 17.1.15 (Unique factorization domain)

An integral domain R is a ***unique factorization domain*** (UFD) if:

(i) every nonzero non-unit $a \in R$ is a product of irreducible elements, $a = p_1 p_2 \cdots p_n$; and

(ii) this factorization is unique up to order and associates: if $a = p_1 \cdots p_n = q_1 \cdots q_m$ with all p_i, q_j irreducible, then $n = m$ and, after reordering, each p_i is an associate of q_i.

Lemma 17.1.16 (Ascending chains in a PID)

In a principal ideal domain, every ascending chain of ideals $I_1 \subseteq I_2 \subseteq \cdots$ is eventually constant.

Proof. The union $I = \bigcup_n I_n$ is an ideal: if $a \in I_m$ and $b \in I_n$, then $a, b \in I_{\max(m,n)}$, so $a + b \in I$, and $ra \in I$ for all $r \in R$. So $I = (d)$ for some d, which lies in some I_N. Then $I = (d) \subseteq I_N \subseteq I_n \subseteq I$ for all $n \geqslant N$. □

Theorem 17.1.17 (Every PID is a UFD)

Every principal ideal domain is a unique factorization domain.

Proof. Existence. Suppose some nonzero non-unit a_0 is not a product of irreducibles. Then a_0 is not irreducible, so $a_0 = bc$ with b, c nonzero non-units, and one of b, c is not a product of irreducibles (otherwise a_0 would be); call it a_1. Then $(a_0) \subsetneq (a_1)$: the inclusion holds since $a_0 = a_1 u$, where u is the other factor, a non-unit; and if $(a_0) = (a_1)$, then $a_0 = a_1 v$ with v a unit, and canceling a_1 would give $u = v$. Repeating the argument with a_1 in place of a_0, and so on, gives a strictly ascending chain $(a_0) \subsetneq (a_1) \subsetneq (a_2) \subsetneq \cdots$, contradicting Lemma 17.1.16.

Uniqueness. We use induction on n. Suppose $p_1 \cdots p_n = q_1 \cdots q_m$ with all factors irreducible. By Theorem 17.1.13, p_n is prime, and since it divides the product $q_1 \cdots q_m$, it divides some q_j (by induction on the number of factors). Reorder so that $j = m$. Then $q_m = p_n u$, and since q_m is irreducible and p_n is not a unit, u is a unit: p_n and q_m are associates. Canceling p_n gives $p_1 \cdots p_{n-1} = u q_1 \cdots q_{m-1}$. If $n = 1$, the left side is 1, and since irreducibles are not units this forces $m = 1$. Likewise, if $m = 1$ then $n = 1$. Otherwise uq_1 is irreducible and an associate of q_1, and the induction hypothesis, applied to $p_1 \cdots p_{n-1} = (uq_1)q_2 \cdots q_{m-1}$, finishes the proof. □

Example 17.1.18 (Failure of unique factorization)

In $R = \mathbb{Z}[\sqrt{-5}]$ we have

$$6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5}).$$

The norm $N(a+b\sqrt{-5}) = |a+b\sqrt{-5}|^2 = a^2+5b^2$ is multiplicative: $N(\alpha\beta) = N(\alpha)N(\beta)$. If α is a unit, then $N(\alpha)N(\alpha^{-1}) = N(1) = 1$, so $N(\alpha) = 1$; and $a^2 + 5b^2 = 1$ only for ± 1. So the units are ± 1, and associates have the same norm. No element has norm 2 or 3, since $a^2 + 5b^2 \in \{2, 3\}$ has no integer solutions.

Now $N(2) = 4$, $N(3) = 9$, and $N(1 \pm \sqrt{-5}) = 6$. If one of these four elements were a product $\alpha\beta$ of two non-units, then $N(\alpha), N(\beta) > 1$ would be proper factors of 4, 9, or 6, so one of them would equal 2 or 3, which is impossible. So all four are irreducible. They are pairwise non-associate, having different norms, except $1 + \sqrt{-5}$ and $1 - \sqrt{-5}$, whose quotient is not ± 1. So the two factorizations of 6 are genuinely different, and R is not a UFD. In particular

2 is irreducible but not prime: $2 \mid (1+\sqrt{-5})(1-\sqrt{-5})$, but 2 divides neither factor, since $\frac{1\pm\sqrt{-5}}{2} \notin R$.

17.1.6 Noetherian Rings

Definition 17.1.19 (Noetherian ring)

A ring R is ***Noetherian*** if it satisfies the ***ascending chain condition*** on ideals: every ascending chain $I_1 \subseteq I_2 \subseteq I_3 \subseteq \cdots$ of ideals of R is eventually constant.

Proposition 17.1.20 (Noetherian means finitely generated ideals)

A ring R is Noetherian if and only if every ideal of R is finitely generated.

Proof. Suppose every ideal is finitely generated, and let $I_1 \subseteq I_2 \subseteq \cdots$ be a chain of ideals. As in the proof of Lemma 17.1.16, the union I is an ideal. It is generated by finitely many elements, which all lie in some I_N; then $I \subseteq I_N$, and $I_n = I_N$ for $n \geqslant N$.

Conversely, suppose some ideal I is not finitely generated. Choose $a_1 \in I$; since $(a_1) \neq I$, choose $a_2 \in I \setminus (a_1)$; since $(a_1, a_2) \neq I$, choose $a_3 \in I \setminus (a_1, a_2)$; and so on. Then $(a_1) \subsetneq (a_1, a_2) \subsetneq (a_1, a_2, a_3) \subsetneq \cdots$ is a chain that is never constant, so R is not Noetherian. □

NOTE Every PID is Noetherian, since every ideal is generated by one element (or by Lemma 17.1.16). In particular fields, $\mathbb{Z}$, and $k[x]$ are Noetherian. The polynomial ring $k[x_1, x_2, \ldots]$ in infinitely many variables is not: $(x_1) \subsetneq (x_1, x_2) \subsetneq (x_1, x_2, x_3) \subsetneq \cdots$.

Theorem 17.1.21 (Hilbert Basis Theorem)

If R is a Noetherian ring, then so is the polynomial ring $R[x]$.

Proof. Let $I \subseteq R[x]$ be an ideal; by Proposition 17.1.20 we must show that I is finitely generated. For $n \geqslant 0$, let $L_n \subseteq R$ be the set of coefficients of x^n in the polynomials $f \in I$ with $f = 0$ or $\deg f \leqslant n$. Since I is an ideal, L_n is an ideal of R; its nonzero elements are the leading coefficients of the elements of I of degree exactly n. If $f \in I$ has degree at most n, then $xf \in I$ has degree at most $n+1$, and its coefficient of x^{n+1} is the coefficient of x^n in f; hence $L_0 \subseteq L_1 \subseteq L_2 \subseteq \cdots$. Since R is Noetherian, there is an N with $L_n = L_N$ for all $n \geqslant N$.

For each $n \leqslant N$, the ideal L_n is finitely generated, so we may choose $f_{n,1}, \ldots, f_{n,k_n} \in I$, each zero or of degree at most n, whose coefficients of x^n generate L_n. Let $J \subseteq I$ be the ideal generated by all the $f_{n,j}$ with $n \leqslant N$. We claim $J = I$. If not, choose $g \in I \setminus J$ of minimal degree d (note $g \neq 0$), with leading coefficient c.

- If $d \leqslant N$, then $c \in L_d$, so $c = \sum_j r_j c_{d,j}$, where $c_{d,j}$ is the coefficient of x^d in $f_{d,j}$. The polynomial $g - \sum_j r_j f_{d,j}$ lies in I, has degree less than d, and is not in J (as $g \notin J$).
- If $d > N$, then $c \in L_d = L_N$, so $c = \sum_j r_j c_{N,j}$, and $g - \sum_j r_j x^{d-N} f_{N,j}$ lies in $I \setminus J$ and has degree less than d.

Either way the minimality of d is contradicted. So $I = J$ is finitely generated. □

Corollary 17.1.22 (Polynomial rings in several variables)

For any Noetherian ring R and $n \geqslant 1$, the ring $R[x_1, \ldots, x_n] = R[x_1, \ldots, x_{n-1}][x_n]$ is Noetherian. In particular $k[x_1, \ldots, x_n]$ is Noetherian for every field k, and so is $\mathbb{Z}[x_1, \ldots, x_n]$.

17.1.7 Polynomials over a Unique Factorization Domain

Gauss's lemma compares factorizations in $\mathbb{Z}[x]$ and in $\mathbb{Q}[x]$. A nonzero polynomial $f \in \mathbb{Z}[x]$ is ***primitive*** if no prime number divides all of its coefficients. Every nonzero $g \in \mathbb{Q}[x]$ can be written as $g = c\, g_0$ with $c \in \mathbb{Q}$ and $g_0 \in \mathbb{Z}[x]$ primitive: multiply g by a common denominator d of its coefficients, and divide dg by the greatest common divisor e of its coefficients, so that $c = e/d$.

Lemma 17.1.23 (Gauss's lemma)

The product of two primitive polynomials in $\mathbb{Z}[x]$ is primitive.

Proof. Let $f, g \in \mathbb{Z}[x]$ be primitive, and suppose a prime p divides every coefficient of fg. Reducing coefficients mod p is a ring homomorphism $\mathbb{Z}[x] \to (\mathbb{Z}/p)[x]$, $h \mapsto \bar{h}$, and $\bar{f}\bar{g} = \overline{fg} = 0$. Since $\mathbb{Z}/p$ is a field, $(\mathbb{Z}/p)[x]$ is an integral domain, so $\bar{f} = 0$ or $\bar{g} = 0$. That is, p divides every coefficient of f or every coefficient of g, a contradiction. □

Proposition 17.1.24 (Factorizations over $\mathbb{Q}$ and over $\mathbb{Z}$)

If $f \in \mathbb{Z}[x]$ is a product $f = gh$ of nonconstant polynomials $g, h \in \mathbb{Q}[x]$, then $f = g_1 h_1$ with $g_1, h_1 \in \mathbb{Z}[x]$, $\deg g_1 = \deg g$, and $\deg h_1 = \deg h$.

Proof. Write $g = a\,g_0$ and $h = b\,h_0$ with $a, b \in \mathbb{Q}$ and $g_0, h_0 \in \mathbb{Z}[x]$ primitive. Then $f = ab\,g_0h_0$, and g_0h_0 is primitive by Gauss's lemma. Write $ab = r/s$ in lowest terms with $s > 0$, so that $sf = r\,g_0h_0$. If a prime p divides s, then p divides every coefficient of $r\,g_0h_0$; as $p \nmid r$, it divides every coefficient of g_0h_0, which is impossible. So $s = 1$, and $f = (rg_0)h_0$. □

Proposition 17.1.25 (Eisenstein's criterion)

Let $f = a_nx^n + \cdots + a_1x + a_0 \in \mathbb{Z}[x]$ with $n \geqslant 1$, and let p be a prime such that $p \nmid a_n$, $p \mid a_i$ for all $i < n$, and $p^2 \nmid a_0$. Then f is irreducible in $\mathbb{Q}[x]$.

Proof. The units of $\mathbb{Q}[x]$ are the nonzero constants, so if f were reducible in $\mathbb{Q}[x]$ it would be a product of two nonconstant polynomials, and by Proposition 17.1.24 we could write $f = gh$ with $g, h \in \mathbb{Z}[x]$ of degrees $r, s \geqslant 1$. The product of the leading coefficients of g and h is a_n, so p divides neither of them, and the reductions mod p satisfy $\deg \bar{g} = r$, $\deg \bar{h} = s$. Now $\bar{g}\bar{h} = \bar{f} = \bar{a}_nx^n$ in $(\mathbb{Z}/p)[x]$. This ring is a Euclidean domain, hence a UFD (Theorems 17.1.12 and 17.1.17), and x is irreducible in it; so by unique factorization $\bar{g} = \beta x^r$ and $\bar{h} = \gamma x^s$ for some $\beta, \gamma \in \mathbb{Z}/p$. Since $r, s \geqslant 1$, the constant terms of $\bar{g}$ and $\bar{h}$ vanish, that is, p divides the constant terms $g(0)$ and $h(0)$. Then p^2 divides $g(0)h(0) = a_0$, a contradiction. □

For instance, $x^3 - 2$, $x^4 - 2$, and $x^5 - 4x + 2$ are irreducible in $\mathbb{Q}[x]$ (take $p = 2$).

The same ideas give a general theorem. For an integral domain R, the ***fraction field*** $\mathrm{Frac}(R)$ is built from R as $\mathbb{Q}$ is built from $\mathbb{Z}$ (and as $k(x)$ is built from $k[x]$ in Chapter 4): its elements are the fractions a/b with $a, b \in R$, $b \neq 0$, where $a/b = a'/b'$ if and only if $ab' = a'b$.

Theorem 17.1.26 (Polynomial rings over a UFD)

If R is a unique factorization domain, then so is $R[x]$. In particular, $\mathbb{Z}[x_1, \ldots, x_n]$ and $k[x_1, \ldots, x_n]$ (for a field k) are unique factorization domains.

Proof (sketch). In a UFD, greatest common divisors exist (up to units), so one can define primitive polynomials in $R[x]$ as above. Every prime element p of R gives an integral domain $R/(p)$ (Proposition 17.1.5), so the proof of Gauss's lemma works with R in place of $\mathbb{Z}$, and so does Proposition 17.1.24, with $K = \mathrm{Frac}(R)$ in place of $\mathbb{Q}$. It follows that the irreducible elements of $R[x]$ are the irreducible elements of R and the primitive polynomials that are irreducible in $K[x]$. Factorizations in $R[x]$ are then obtained, and shown to be unique, from those in R and in $K[x]$,

which is a UFD because it is a Euclidean domain. The second statement follows by induction on n, since $R[x_1, \dots, x_n] = R[x_1, \dots, x_{n-1}][x_n]$. □

NOTE We have shown the implications

$$\text{Euclidean domain} \implies \text{PID} \implies \text{UFD} \implies \text{integral domain},$$

and none of them reverses. The ring $\mathbb{Z}\left[\frac{1+\sqrt{-19}}{2}\right]$ is a PID but not Euclidean. The ring $\mathbb{Z}[x]$ is a UFD by Theorem 17.1.26 but not a PID: if $(2, x) = (f)$, then $f \mid 2$ forces $f = \pm 1$ or ± 2, and $f \mid x$ forces $f = \pm 1$; but $1 \notin (2, x)$, since every element $2g + xh$ of $(2, x)$ has an even constant term. Finally $\mathbb{Z}[\sqrt{-5}]$ is an integral domain but not a UFD (Example 17.1.18).

Exercise 17.1.27 (Exercises)

(1) Show that $\mathbb{Z}[\sqrt{-5}]$ is not a PID by finding an ideal that is not principal. (Hint: consider $(2, 1 + \sqrt{-5})$.)
(2) Prove that $\mathbb{Z}[i]$ is a Euclidean domain with $N(a + bi) = a^2 + b^2$. (The key step: for any $\alpha, \beta \in \mathbb{Z}[i]$ with $\beta \neq 0$, the quotient $\alpha/\beta \in \mathbb{C}$ lies within distance $\frac{\sqrt{2}}{2} < 1$ of some Gaussian integer.)

Solution. **(1)** Let $R = \mathbb{Z}[\sqrt{-5}]$ and $I = (2, 1 + \sqrt{-5})$. First, $I = \{a + b\sqrt{-5} : a \equiv b \pmod 2\}$. Indeed, the right side is an additive subgroup containing 2 and $1 + \sqrt{-5}$, and it is closed under multiplication by $\sqrt{-5}$, since $\sqrt{-5}(a + b\sqrt{-5}) = -5b + a\sqrt{-5}$ and $-5b \equiv b \equiv a \pmod 2$; so it is an ideal containing I. Conversely, if $a \equiv b \pmod 2$, then $a + b\sqrt{-5} = b(1 + \sqrt{-5}) + (a - b)$ with $a - b \in (2)$, so it lies in I. In particular $1 \notin I$, and $I \neq R$.
Suppose $I = (\alpha)$. Then $\alpha \mid 2$ and $\alpha \mid 1 + \sqrt{-5}$, so by multiplicativity of the norm (Example 17.1.18), $N(\alpha)$ divides $N(2) = 4$ and $N(1 + \sqrt{-5}) = 6$, hence divides 2. No element has norm 2, so $N(\alpha) = 1$, $\alpha = \pm 1$, and $I = R$, a contradiction. So I is not principal.
(2) $\mathbb{Z}[i]$ is an integral domain, being a subring of $\mathbb{C}$. Given $\alpha, \beta \in \mathbb{Z}[i]$ with $\beta \neq 0$, write $\alpha/\beta = u + vi$ with $u, v \in \mathbb{Q}$. Choose integers m, n with $|u - m| \leqslant \frac{1}{2}$ and $|v - n| \leqslant \frac{1}{2}$, and set $\gamma = m + ni$ and $r = \alpha - \beta\gamma \in \mathbb{Z}[i]$. Since $N(z) = |z|^2$ is multiplicative on $\mathbb{C}$,

$$N(r) = N(\beta)\left|\frac{\alpha}{\beta} - \gamma\right|^2 = N(\beta)\big((u - m)^2 + (v - n)^2\big) \leqslant \tfrac{1}{2}N(\beta) < N(\beta).$$

So $\alpha = \beta\gamma + r$ with $r = 0$ or $N(r) < N(\beta)$.

17.2 Galois Theory

17.2.1 Field Extensions

Definition 17.2.1 (Field extension and degree)

A ***field extension*** L/K is a field L containing a field K as a subfield (Chapter 4). Then L is a vector space over K, and the ***degree*** of the extension is $[L : K] = \dim_K L$. The extension is ***finite*** if $[L : K] < \infty$.

Proposition 17.2.2 (Tower law)

If $K \subseteq L \subseteq M$ are fields and $[M : L]$ and $[L : K]$ are finite, then $[M : K] = [M : L] \cdot [L : K]$.

Proof. Let $e_1, \dots, e_m$ be a basis of M over L and $f_1, \dots, f_n$ a basis of L over K. We show that the products $f_j e_i$ form a basis of M over K. Every $x \in M$ is $x = \sum_i a_i e_i$ with $a_i \in L$, and each $a_i = \sum_j b_{ij} f_j$ with $b_{ij} \in K$; so $x = \sum_{i,j} b_{ij} f_j e_i$. If $\sum_{i,j} b_{ij} f_j e_i = 0$ with $b_{ij} \in K$, then $\sum_i \big(\sum_j b_{ij} f_j\big) e_i = 0$, so each $\sum_j b_{ij} f_j = 0$ by independence of the e_i over L, and then each $b_{ij} = 0$ by independence of the f_j over K. $\square$

Definition 17.2.3 (Algebraic and transcendental elements)

Let L/K be a field extension and $\alpha \in L$. Then α is ***algebraic*** over K if $f(\alpha) = 0$ for some nonzero $f \in K[x]$, and ***transcendental*** over K otherwise.

For $\alpha \in L$, evaluation $\mathrm{ev}_\alpha : K[x] \to L$, $f \mapsto f(\alpha)$, is a ring homomorphism, and α is algebraic exactly when its kernel is nonzero. For $\alpha_1, \dots, \alpha_n \in L$, we write $K(\alpha_1, \dots, \alpha_n)$ for the smallest subfield of L containing K and $\alpha_1, \dots, \alpha_n$ (the intersection of all such subfields).

Proposition 17.2.4 (Minimal polynomial)

Let $\alpha \in L$ be algebraic over K. There is a unique monic polynomial $m_\alpha \in K[x]$ with $\ker(\mathrm{ev}_\alpha) = (m_\alpha)$, the ***minimal polynomial*** of α over K. It is the monic polynomial of least degree with $m_\alpha(\alpha) = 0$, and it is irreducible.

Proof. The kernel of ev_α is a nonzero ideal of $K[x]$, which is a PID (Example 17.1.10 and Theorem 17.1.12); a generator is a nonzero element of least degree, and exactly one generator is monic, since two monic polynomials that divide each other are equal. If $m_\alpha = fg$ with $\deg f, \deg g < \deg m_\alpha$, then $f(\alpha)g(\alpha) = 0$ in the

field L, so $f(\alpha) = 0$ or $g(\alpha) = 0$, contradicting minimality. And m_α is not a unit, since $\deg m_\alpha \geqslant 1$. So m_α is irreducible. $\square$

Proposition 17.2.5 (Simple algebraic extensions)

If α is algebraic over K with minimal polynomial m_α of degree n, then

$$K(\alpha) = \{f(\alpha) : f \in K[x]\} \cong K[x]/(m_\alpha),$$

and $[K(\alpha) : K] = n$, with basis $1, \alpha, \alpha^2, \ldots, \alpha^{n-1}$.

Proof. By Proposition 17.1.2, the image $K[\alpha] = \{f(\alpha) : f \in K[x]\}$ of ev_α is isomorphic to $K[x]/(m_\alpha)$. Since m_α is irreducible, (m_α) is maximal (Corollary 17.1.14), so $K[\alpha]$ is a field. It contains K and α, and it is contained in every subfield of L containing them; so $K[\alpha] = K(\alpha)$. Division with remainder by m_α shows that every element of $K(\alpha)$ is $f(\alpha)$ with $f = 0$ or $\deg f < n$, and this f is unique: if $f(\alpha) = g(\alpha)$ with $\deg f, \deg g < n$, then $m_\alpha \mid f - g$, so $f = g$. Hence $1, \alpha, \ldots, \alpha^{n-1}$ is a basis. $\square$

Example 17.2.6 (Two simple extensions of $\mathbb{Q}$)

The minimal polynomial of $\sqrt{2}$ over $\mathbb{Q}$ is $x^2 - 2$, so $\mathbb{Q}(\sqrt{2}) \cong \mathbb{Q}[x]/(x^2 - 2)$ has degree 2 over $\mathbb{Q}$, with basis $1, \sqrt{2}$. The minimal polynomial of $\sqrt[3]{2}$ over $\mathbb{Q}$ is $x^3 - 2$, which is irreducible by Eisenstein's criterion (Proposition 17.1.25) with $p = 2$; so $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3$.

17.2.2 Splitting Fields and Separability

Definition 17.2.7 (Splitting field)

Let $f \in K[x]$ be nonconstant. A ***splitting field*** of f over K is an extension L/K such that f factors into linear factors in $L[x]$ and $L = K(\alpha_1, \ldots, \alpha_n)$, where $\alpha_1, \ldots, \alpha_n$ are the roots of f in L.

Theorem 17.2.8 (Existence and uniqueness of splitting fields)

Every nonconstant $f \in K[x]$ has a splitting field over K. Any two splitting fields of f over K are isomorphic by an isomorphism that is the identity on K.

Proof of existence. We use induction on $\deg f$. Let q be an irreducible factor of f in $K[x]$, which exists because $K[x]$ is a UFD. By Corollary 17.1.14, $K_1 = K[x]/(q)$ is a field. The composite $K \to K[x] \to K_1$ is a field homomorphism, hence injective (Proposition 4.2.3), so we may regard K_1 as an extension of K; and the class α of

x satisfies $q(\alpha) = 0$, hence $f(\alpha) = 0$. In $K_1[x]$ we can write $f = (x - \alpha)f_1$. If f_1 is constant, $K_1 = K(\alpha)$ is a splitting field. Otherwise, by induction f_1 has a splitting field L over K_1. Then $f = (x - \alpha)f_1$ factors into linear factors in $L[x]$, and L is generated over K by α and the roots of f_1, because $K_1 = K(\alpha)$; so L is a splitting field of f over K. □

We omit the proof of uniqueness, which extends an isomorphism of fields one root at a time; see Artin, *Algebra* (2nd edition), Chapters 15 and 16.

Definition 17.2.9 (Normal extension)

A finite extension L/K is ***normal*** if every irreducible polynomial in $K[x]$ that has a root in L factors into linear factors in $L[x]$.

Definition 17.2.10 (Separable polynomials and extensions)

A nonconstant polynomial $f \in K[x]$ is ***separable*** if it has no repeated roots in a splitting field. An algebraic element α of an extension of K is separable over K if its minimal polynomial is separable, and an extension L/K is ***separable*** if every element of L is algebraic and separable over K.

The ***formal derivative*** of $f = \sum_i a_i x^i$ is $f' = \sum_i i a_i x^{i-1}$. The product rule $(fg)' = f'g + fg'$ holds: it is easily checked for monomials and extends by bilinearity.

Proposition 17.2.11 (Separability in characteristic 0)

If $\operatorname{char} K = 0$, every irreducible polynomial in $K[x]$ is separable. Hence every algebraic extension of K is separable.

Proof. If α is a repeated root of f in some extension L, then $f = (x - \alpha)^2 g$ in $L[x]$, so $f' = 2(x - \alpha)g + (x - \alpha)^2 g'$ and $f'(\alpha) = 0$. Now let $f \in K[x]$ be irreducible of degree $n \geqslant 1$, with leading coefficient a_n. Since $\operatorname{char} K = 0$, the coefficient $n a_n$ of x^{n-1} in f' is nonzero, so $f' \neq 0$ and $\deg f' < n$. The ideal (f, f') of the PID $K[x]$ is (d) for some $d \mid f$; as f does not divide f', d is not an associate of f, so d is a unit because f is irreducible. Thus $1 = af + bf'$ for some $a, b \in K[x]$. If α were a repeated root of f, evaluating at α would give $1 = 0$. The second statement follows because minimal polynomials are irreducible (Proposition 17.2.4). □

17.2.3 The Galois Group

For an extension L/K, let $\mathrm{Aut}_K(L)$ be the group of field automorphisms $\sigma : L \to L$ with $\sigma|_K = \mathrm{id}_K$.

Lemma 17.2.12 (Automorphisms permute roots)

Let L/K be an extension and $f \in K[x]$.

(i) If $\sigma \in \mathrm{Aut}_K(L)$ and $f(\alpha) = 0$, then $f(\sigma(\alpha)) = 0$.

(ii) If $L = K(\alpha_1, \ldots, \alpha_n)$, where $\alpha_1, \ldots, \alpha_n$ are the distinct roots of f in L, then sending σ to the permutation of $\{\alpha_1, \ldots, \alpha_n\}$ that it induces is an injective homomorphism $\mathrm{Aut}_K(L) \to S_n$.

Proof. (i) Since σ is a ring homomorphism fixing the coefficients of f, we have $f(\sigma(\alpha)) = \sigma(f(\alpha)) = 0$.

(ii) By (i), σ maps the finite set $\{\alpha_1, \ldots, \alpha_n\}$ injectively to itself, hence permutes it, and composition of automorphisms corresponds to composition of permutations. If σ fixes every α_i, then $\{x \in L : \sigma(x) = x\}$ is a subfield of L containing K and all the α_i, so it is L, and $\sigma = \mathrm{id}$. □

The following theorem characterizes the extensions with as many automorphisms as possible. We state it without proof; see Artin, *Algebra*, Chapter 16.

Theorem 17.2.13 (Galois extensions)

Let L/K be a finite extension. Then $|\mathrm{Aut}_K(L)| \leqslant [L : K]$, and the following are equivalent:

(i) $|\mathrm{Aut}_K(L)| = [L : K]$;

(ii) L/K is normal and separable;

(iii) L is a splitting field over K of a separable polynomial;

(iv) the elements of L fixed by every element of $\mathrm{Aut}_K(L)$ are exactly the elements of K.

Moreover, L/K is normal if and only if L is a splitting field over K of some polynomial.

Definition 17.2.14 (Galois extension and Galois group)

A finite extension L/K satisfying the equivalent conditions of Theorem 17.2.13 is ***Galois***, and its ***Galois group*** is $\mathrm{Gal}(L/K) = \mathrm{Aut}_K(L)$. Thus $|\mathrm{Gal}(L/K)| = [L : K]$.

In characteristic 0 the splitting field $L = K(\alpha_1, \ldots, \alpha_n)$ of any nonconstant polynomial is a Galois extension of K: it is finite by Proposition 17.2.5 and the tower law (each α_i is algebraic over $K(\alpha_1, \ldots, \alpha_{i-1})$), normal by the last statement of Theorem 17.2.13, and separable by Proposition 17.2.11.

Example 17.2.15 (Quadratic extensions)
Let $d \in \mathbb{Q}$ be a non-square. Then $\mathbb{Q}(\sqrt{d})$ is the splitting field of $x^2 - d$ over $\mathbb{Q}$, so it is Galois of degree 2. An automorphism must send $\sqrt{d}$ to a root of $x^2 - d$, that is, to $\pm\sqrt{d}$, so $\mathrm{Gal}(\mathbb{Q}(\sqrt{d})/\mathbb{Q}) = \{\mathrm{id}, \sigma\}$ with $\sigma(\sqrt{d}) = -\sqrt{d}$, and $\mathrm{Gal}(\mathbb{Q}(\sqrt{d})/\mathbb{Q}) \cong \mathbb{Z}/2$.

Example 17.2.16 (The splitting field of $x^3 - 2$)
Let $\omega = e^{2\pi i/3}$. The roots of $x^3 - 2$ in $\mathbb{C}$ are

$$\alpha_1 = \sqrt[3]{2}, \qquad \alpha_2 = \omega\sqrt[3]{2}, \qquad \alpha_3 = \omega^2\sqrt[3]{2},$$

and they generate $L = \mathbb{Q}(\sqrt[3]{2}, \omega)$, since $\omega = \alpha_2/\alpha_1$; so L is the splitting field of $x^3 - 2$ over $\mathbb{Q}$, and $L/\mathbb{Q}$ is Galois. We have $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3$, and ω is a root of $x^2 + x + 1$ but does not lie in $\mathbb{Q}(\sqrt[3]{2}) \subset \mathbb{R}$, so $[L : \mathbb{Q}(\sqrt[3]{2})] = 2$ and $[L : \mathbb{Q}] = 6$ by the tower law. By Lemma 17.2.12, $\mathrm{Gal}(L/\mathbb{Q})$ embeds in S_3 through its action on $\alpha_1, \alpha_2, \alpha_3$; since both groups have order 6, $\mathrm{Gal}(L/\mathbb{Q}) \cong S_3$.

17.2.4 The Fundamental Theorem of Galois Theory

Theorem 17.2.17 (Fundamental theorem of Galois theory)
Let L/K be a finite Galois extension with Galois group $G = \mathrm{Gal}(L/K)$. There is an inclusion-reversing bijection

$$\{\text{intermediate fields } K \subseteq E \subseteq L\} \longleftrightarrow \{\text{subgroups } H \leqslant G\},$$

given by $E \mapsto \mathrm{Gal}(L/E)$ and $H \mapsto L^H = \{x \in L : \sigma(x) = x \text{ for all } \sigma \in H\}$. If E corresponds to H, then:

(i) $[L : E] = |H|$ and $[E : K] = [G : H]$;

(ii) E/K is Galois if and only if H is a normal subgroup of G, and in that case $\mathrm{Gal}(E/K) \cong G/H$;

(iii) the intersection of two intermediate fields corresponds to the subgroup generated by the two subgroups, and the smallest field containing both corresponds to the intersection of the subgroups.

NOTE We do not prove the theorem. Besides Theorem 17.2.13, the proof uses the primitive element theorem (every finite separable extension is of the form $K(\alpha)$) and Artin's lemma ($[L : L^H] = |H|$ for every finite group H of automorphisms of a field L).

Example 17.2.18 (Galois correspondence for $x^3 - 2$)

Let $L = \mathbb{Q}(\sqrt[3]{2}, \omega)$ as in Example 17.2.16, and identify $G = \mathrm{Gal}(L/\mathbb{Q})$ with S_3 through the numbering $\alpha_1, \alpha_2, \alpha_3$ of the roots.

- The subgroup $A_3 = \langle (123) \rangle$ has index 2 and is normal. The 3-cycle $\sigma = (123)$ sends $\alpha_1 \mapsto \alpha_2 \mapsto \alpha_3$, so $\sigma(\omega) = \sigma(\alpha_2/\alpha_1) = \alpha_3/\alpha_2 = \omega$; hence A_3 fixes $\mathbb{Q}(\omega)$, which has degree $2 = [G : A_3]$, and A_3 corresponds to $\mathbb{Q}(\omega)$. Accordingly $\mathbb{Q}(\omega)/\mathbb{Q}$ is Galois with group $S_3/A_3 \cong \mathbb{Z}/2$.
- The subgroup $\langle (23) \rangle$ fixes α_1, so it corresponds to $\mathbb{Q}(\alpha_1) = \mathbb{Q}(\sqrt[3]{2})$, which has degree $3 = [G : \langle (23) \rangle]$. Likewise $\langle (13) \rangle$ corresponds to $\mathbb{Q}(\omega\sqrt[3]{2})$ and $\langle (12) \rangle$ to $\mathbb{Q}(\omega^2\sqrt[3]{2})$. These subgroups are not normal, and the three cubic fields are not Galois over $\mathbb{Q}$; for instance $x^3 - 2$ has the root $\sqrt[3]{2}$ in $\mathbb{Q}(\sqrt[3]{2})$ but does not split there.

17.2.5 Cyclotomic Extensions

Let $n \geqslant 1$ and $\zeta_n = e^{2\pi i/n}$. The nth roots of unity are ζ_n^a for $a \in \mathbb{Z}/n$, and ζ_n^a is a ***primitive*** nth root of unity (one of order exactly n) if and only if $a \in (\mathbb{Z}/n)^\times$. The nth ***cyclotomic polynomial*** is

$$\Phi_n(x) = \prod_{a \in (\mathbb{Z}/n)^\times} (x - \zeta_n^a),$$

of degree $\varphi(n) = |(\mathbb{Z}/n)^\times|$, Euler's totient. The field $\mathbb{Q}(\zeta_n)$ is the nth ***cyclotomic field***. We state the following theorem without proof.

Theorem 17.2.19 (Irreducibility of cyclotomic polynomials)

The polynomial Φ_n has integer coefficients and is irreducible in $\mathbb{Q}[x]$. Hence it is the minimal polynomial of ζ_n over $\mathbb{Q}$, and $[\mathbb{Q}(\zeta_n) : \mathbb{Q}] = \varphi(n)$.

Proposition 17.2.20 (Galois group of a cyclotomic field)

The extension $\mathbb{Q}(\zeta_n)/\mathbb{Q}$ is Galois, and $\mathrm{Gal}(\mathbb{Q}(\zeta_n)/\mathbb{Q}) \cong (\mathbb{Z}/n)^\times$: the class of a corresponds to the automorphism σ_a with $\sigma_a(\zeta_n) = \zeta_n^a$.

Proof. The roots of $x^n - 1$ are the powers of ζ_n, so $\mathbb{Q}(\zeta_n)$ is the splitting field of $x^n - 1$ and is Galois over $\mathbb{Q}$, of degree $\varphi(n)$ by Theorem 17.2.19. An automorphism σ sends ζ_n to a root of its minimal polynomial Φ_n (Lemma 17.2.12), so $\sigma(\zeta_n) = \zeta_n^a$ for a unique $a \in (\mathbb{Z}/n)^\times$, and σ is determined by a because ζ_n generates the field. Since $\sigma_a\sigma_b(\zeta_n) = \sigma_a(\zeta_n^b) = \zeta_n^{ab}$, the map $\sigma \mapsto a$ is an injective homomorphism $\mathrm{Gal}(\mathbb{Q}(\zeta_n)/\mathbb{Q}) \to (\mathbb{Z}/n)^\times$. Both groups have order $\varphi(n)$, so it is an isomorphism. □

NOTE Since $(\mathbb{Z}/n)^\times$ is abelian, every subgroup of it is normal, so by the fundamental theorem every intermediate field of $\mathbb{Q}(\zeta_n)/\mathbb{Q}$ is Galois over $\mathbb{Q}$ with abelian Galois group. The Kronecker–Weber theorem, a deep result of algebraic number theory, says conversely that every finite Galois extension of $\mathbb{Q}$ with abelian Galois group is contained in some cyclotomic field.

17.2.6 Solvability by Radicals

Definition 17.2.21 (Solvable group)

A group G is ***solvable*** if it has a chain of subgroups

$$G = G_0 \supseteq G_1 \supseteq \cdots \supseteq G_m = \{e\}$$

in which each G_{i+1} is normal in G_i and each quotient G_i/G_{i+1} is abelian.

Proposition 17.2.22 (Subgroups of solvable groups)

Every subgroup of a solvable group is solvable.

Proof. Let $G = G_0 \supseteq \cdots \supseteq G_m = \{e\}$ be as in the definition, and $H \leqslant G$. The homomorphism $H \cap G_i \to G_i/G_{i+1}$, $x \mapsto xG_{i+1}$, has kernel $H \cap G_{i+1}$, which is therefore normal in $H \cap G_i$, and by the First Isomorphism Theorem (Theorem 3.3.3) the quotient $(H \cap G_i)/(H \cap G_{i+1})$ is isomorphic to a subgroup of the abelian group G_i/G_{i+1}. So the chain $H \cap G_i$ shows that H is solvable. □

Example 17.2.23 (Symmetric groups)

- For $n \leqslant 4$, S_n is solvable. The groups S_1 and S_2 are abelian; $S_3 \supseteq A_3 \supseteq \{e\}$ has quotients $\mathbb{Z}/2$ and $\mathbb{Z}/3$; and $S_4 \supseteq A_4 \supseteq K \supseteq \{e\}$, with $K = \{e, (12)(34), (13)(24), (14)(23)\}$, has quotients of orders $2, 3, 4$, namely $\mathbb{Z}/2$, $\mathbb{Z}/3$, and $K \cong \mathbb{Z}/2 \times \mathbb{Z}/2$. Each subgroup is normal in the previous

one: A_n is the kernel of the sign, and K consists of e and a full conjugacy class of S_4.

- S_5 is not solvable. If N is a normal subgroup of a group G' with G'/N abelian, then N contains the commutator subgroup $[G', G']$ (Definition 3.3.9), since every commutator $aba^{-1}b^{-1}$ maps to the identity of the abelian group G'/N. The commutator subgroup $[A_5, A_5]$ is normal in A_5 and nontrivial, as A_5 is not abelian; since A_5 is simple (Theorem 11.4.9), $[A_5, A_5] = A_5$. Now let $S_5 = G_0 \supseteq G_1 \supseteq \cdots$ be a chain as in the definition. Then every G_i contains A_5: this holds for G_0, and if $G_i \supseteq A_5$, then $G_{i+1} \supseteq [G_i, G_i] \supseteq [A_5, A_5] = A_5$. So the chain never reaches $\{e\}$.

Definition 17.2.24 (Radical extension)

An extension L/K is ***radical*** if there is a chain of fields $K = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_m = L$ with $K_{i+1} = K_i(\alpha_i)$ and $\alpha_i^{n_i} \in K_i$ for some $n_i \geqslant 1$. A polynomial $f \in K[x]$ is ***solvable by radicals*** if its splitting field over K is contained in a radical extension of K.

So f is solvable by radicals when its roots can be obtained from elements of K by field operations and the extraction of nth roots. The following theorem, which we state without proof (see Artin, *Algebra*, Chapter 16), translates this into group theory.

Theorem 17.2.25 (Galois's criterion)

Let K be a field of characteristic 0 and $f \in K[x]$ nonconstant, with splitting field L. Then f is solvable by radicals if and only if $\mathrm{Gal}(L/K)$ is a solvable group.

Lemma 17.2.26 (Generating S_p)

Let p be a prime. A transposition and a p-cycle together generate S_p.

Proof. Let $\tau = (a\ b)$ be a transposition and σ a p-cycle. Since σ moves every point through a single cycle, $\sigma^k(a) = b$ for some $1 \leqslant k < p$. The power σ^k has order p (as p is prime), so it is again a p-cycle. Renumber the points so that $\sigma^k = (1\,2 \cdots p)$, $a = 1$, and $b = 2$. Conjugating $\tau = (1\,2)$ by the powers of σ^k gives $\sigma^{kj}(1\,2)\sigma^{-kj} = (1+j\ \ 2+j)$, so the subgroup generated contains all adjacent transpositions $(i\ \ i+1)$, which generate S_p (Proposition 3.4.9). □

Example 17.2.27 (A quintic with Galois group S_5)

Let $f = x^5 - 4x + 2 \in \mathbb{Q}[x]$. It is irreducible by Eisenstein's criterion with $p = 2$. Let $L \subset \mathbb{C}$ be its splitting field and $G = \mathrm{Gal}(L/\mathbb{Q})$. The five roots of f are distinct (Proposition 17.2.11), and by Lemma 17.2.12 G is isomorphic to a subgroup of S_5.

- Since $f(-2) = -22$, $f(0) = 2$, $f(1) = -1$, and $f(2) = 26$, the polynomial f has at least three real roots. Its derivative $5x^4 - 4$ has only two real zeros, so by Rolle's theorem f has at most three real roots. So f has exactly three real roots and two nonreal, complex conjugate roots.
- Complex conjugation permutes the roots of f, which generate L, so it maps L to itself and restricts to an element of G. It fixes the three real roots and swaps the other two: it is a transposition.
- For a root α, $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 5$ divides $[L : \mathbb{Q}] = |G|$ by the tower law. So G contains an element of order 5, by Cauchy's theorem (Corollary 11.5.4); and the elements of order 5 in S_5 are the 5-cycles.

By Lemma 17.2.26, $G \cong S_5$.

Corollary 17.2.28 (The quintic is not solvable by radicals)

The polynomial $x^5 - 4x + 2$ is not solvable by radicals over $\mathbb{Q}$. Consequently there is no formula expressing the roots of a general polynomial of degree 5 in terms of its coefficients using only the field operations and radicals.

Proof. By Example 17.2.27, its Galois group is S_5, which is not solvable (Example 17.2.23); apply Theorem 17.2.25. □

NOTE The impossibility of a general formula is the Abel–Ruffini theorem (Abel, 1824); the group-theoretic criterion is due to Galois (1832). For polynomials of degree at most 4, the Galois group is a subgroup of S_n with $n \leqslant 4$, hence solvable (Proposition 17.2.22), which is why the quadratic formula and the formulas of Cardano and Ferrari exist.

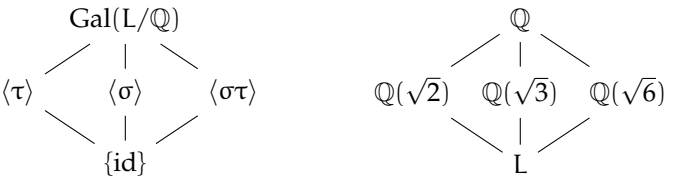
Exercise 17.2.29 (Exercises)

(1) Compute $\mathrm{Gal}(\mathbb{Q}(\sqrt{2}, \sqrt{3})/\mathbb{Q})$, and draw its lattice of subgroups alongside the lattice of intermediate fields.

(2) Show that the Galois group of $x^4 - 2$ over $\mathbb{Q}$ is isomorphic to the dihedral group D_4 of order 8 (Definition 2.2.14).

Solution. **(1)** Let $L = \mathbb{Q}(\sqrt{2}, \sqrt{3})$. First, $\sqrt{3} \notin \mathbb{Q}(\sqrt{2})$: if $\sqrt{3} = a + b\sqrt{2}$ with $a, b \in \mathbb{Q}$, then $3 = a^2 + 2b^2 + 2ab\sqrt{2}$, so $ab = 0$; but $b = 0$ would make $\sqrt{3}$ rational, and $a = 0$ would give $b^2 = 3/2$. So $x^2 - 3$ is irreducible over $\mathbb{Q}(\sqrt{2})$, and $[L : \mathbb{Q}] = 2 \cdot 2 = 4$ by the tower law. The field L is the splitting field of $(x^2-2)(x^2-3)$, so it is Galois over $\mathbb{Q}$ and $|\mathrm{Gal}(L/\mathbb{Q})| = 4$. An automorphism sends $\sqrt{2} \mapsto \pm\sqrt{2}$ and $\sqrt{3} \mapsto \pm\sqrt{3}$ and is determined by these signs; as there are four automorphisms, all four sign choices occur. Let σ change the sign of $\sqrt{2}$ only and τ that of $\sqrt{3}$ only. Then $\sigma^2 = \tau^2 = \mathrm{id}$ and $\sigma\tau = \tau\sigma$, so $\mathrm{Gal}(L/\mathbb{Q}) = \{\mathrm{id}, \sigma, \tau, \sigma\tau\} \cong \mathbb{Z}/2 \times \mathbb{Z}/2$.

The three subgroups of order 2 are $\langle\tau\rangle$, $\langle\sigma\rangle$, $\langle\sigma\tau\rangle$; they fix $\sqrt{2}$, $\sqrt{3}$, and $\sqrt{6} = \sqrt{2}\sqrt{3}$ respectively, so they correspond to $\mathbb{Q}(\sqrt{2})$, $\mathbb{Q}(\sqrt{3})$, $\mathbb{Q}(\sqrt{6})$. The two lattices, with corresponding positions matched (the correspondence reverses inclusions):



(In the field lattice, larger fields are drawn lower, so that the two pictures correspond position by position.)

(2) The polynomial $x^4 - 2$ is irreducible by Eisenstein's criterion with $p = 2$, and its roots are $\pm\sqrt[4]{2}$ and $\pm i\sqrt[4]{2}$, so its splitting field is $L = \mathbb{Q}(\sqrt[4]{2}, i)$. We have $[\mathbb{Q}(\sqrt[4]{2}) : \mathbb{Q}] = 4$, and $i \notin \mathbb{Q}(\sqrt[4]{2}) \subset \mathbb{R}$ is a root of $x^2 + 1$, so $[L : \mathbb{Q}] = 8$ and $G = \mathrm{Gal}(L/\mathbb{Q})$ has order 8. An automorphism sends $\sqrt[4]{2}$ to one of the four roots $i^a\sqrt[4]{2}$ and i to $\pm i$, and is determined by these choices; since there are eight automorphisms, all eight choices occur. Let $\sigma(\sqrt[4]{2}) = i\sqrt[4]{2}$, $\sigma(i) = i$, and let τ be complex conjugation: $\tau(\sqrt[4]{2}) = \sqrt[4]{2}$, $\tau(i) = -i$. Then σ has order 4, τ has order 2, and $\tau\sigma\tau = \sigma^{-1}$: both sides send $i \mapsto i$ and $\sqrt[4]{2} \mapsto -i\sqrt[4]{2}$. Since $\tau \notin \langle\sigma\rangle$ (it moves i), $G = \langle\sigma\rangle \cup \tau\langle\sigma\rangle$, and $\sigma^a\tau^b \mapsto r^a s^b$ is an isomorphism onto $D_4 = \langle r, s \mid r^4 = s^2 = e,\ srs^{-1} = r^{-1}\rangle$. Acting on the four roots, G is a subgroup of S_4 of order 8, not all of S_4.

17.3 Lie Groups and Lie Algebras

17.3.1 Matrix Lie Groups

The finite groups studied so far (symmetric groups, dihedral groups, and so on) are discrete. Many important groups in mathematics and physics are continuous:

the rotation group $\mathrm{SO}(3)$, the group $\mathrm{GL}_n(\mathbb{R})$ of invertible matrices, and others. Lie theory studies such groups through linear algebra. We give $\mathbb{C}^{n\times n} \cong \mathbb{C}^{n^2}$ its usual topology, in which $A_m \to A$ when every entry converges.

Definition 17.3.1 (Matrix Lie group)

A ***matrix Lie group*** is a subgroup $G \subseteq \mathrm{GL}_n(\mathbb{C})$ that is closed in $\mathrm{GL}_n(\mathbb{C})$: if $A_m \in G$ and $A_m \to A \in \mathrm{GL}_n(\mathbb{C})$, then $A \in G$.

A theorem of von Neumann and Cartan, which we do not prove, says that a matrix Lie group is a smooth submanifold of $\mathbb{C}^{n\times n}$ on which multiplication and inversion are smooth maps; that is, it is a ***Lie group***.

Example 17.3.2 (The classical matrix Lie groups)

Each of the following is closed in $\mathrm{GL}_n(\mathbb{C})$, being defined by equations in the matrix entries and their complex conjugates (see also Chapter 9).

- $\mathrm{GL}_n(\mathbb{R})$ and $\mathrm{GL}_n(\mathbb{C})$, the general linear groups.
- $\mathrm{SL}_n(\mathbb{R}) = \{A \in \mathrm{GL}_n(\mathbb{R}) : \det A = 1\}$ and $\mathrm{SL}_n(\mathbb{C})$, the special linear groups.
- $\mathrm{O}(n) = \{A \in \mathrm{GL}_n(\mathbb{R}) : A^{t}A = I\}$, the orthogonal group, and $\mathrm{SO}(n) = \mathrm{O}(n) \cap \mathrm{SL}_n(\mathbb{R})$, the group of rotations.
- $\mathrm{U}(n) = \{A \in \mathrm{GL}_n(\mathbb{C}) : A^{*}A = I\}$, the unitary group, and $\mathrm{SU}(n) = \mathrm{U}(n) \cap \mathrm{SL}_n(\mathbb{C})$.
- $\mathrm{Sp}(2n,\mathbb{R}) = \{A \in \mathrm{GL}_{2n}(\mathbb{R}) : A^{t}JA = J\}$, where $J = \left(\begin{smallmatrix} 0 & I_n \\ -I_n & 0 \end{smallmatrix}\right)$, the symplectic group (Section 9.7).

NOTE The dimension of a Lie group as a manifold can be guessed by counting equations: $\mathrm{GL}_n(\mathbb{R})$ has dimension n^2; the single equation $\det A = 1$ gives $\dim \mathrm{SL}_n(\mathbb{R}) = n^2 - 1$; and $A^{t}A = I$ consists of $\frac{n(n+1)}{2}$ equations, since $A^{t}A$ is symmetric, giving $\dim \mathrm{O}(n) = n^2 - \frac{n(n+1)}{2} = \frac{n(n-1)}{2}$. The Lie algebras below confirm these counts.

17.3.2 The Matrix Exponential

Definition 17.3.3 (Matrix exponential)

For $X \in \mathbb{C}^{n\times n}$, the ***matrix exponential*** is

$$e^X = \sum_{k=0}^{\infty} \frac{X^k}{k!} = I + X + \frac{X^2}{2!} + \frac{X^3}{3!} + \cdots .$$

The series converges absolutely for every X: for the operator norm $\|X\| = \max\{|Xv| : v \in \mathbb{C}^n,\ |v| = 1\}$, where $|v|$ is the standard length on $\mathbb{C}^n$, we have $\|XY\| \leqslant \|X\|\,\|Y\|$, hence $\|X^k\| \leqslant \|X\|^k$, so the norms of the terms are bounded by those of the series for $e^{\|X\|}$.

Proposition 17.3.4 (Properties of the matrix exponential)

Let $X, Y \in \mathbb{C}^{n\times n}$.

(i) $e^0 = I$.

(ii) If $XY = YX$, then $e^{X+Y} = e^X e^Y$. In particular e^X is invertible with inverse e^{-X}, and $t \mapsto e^{tX}$ is a homomorphism $(\mathbb{R}, +) \to \mathrm{GL}_n(\mathbb{C})$.

(iii) $e^{AXA^{-1}} = Ae^XA^{-1}$ for $A \in \mathrm{GL}_n(\mathbb{C})$.

(iv) $\det(e^X) = e^{\operatorname{Tr} X}$.

(v) $\frac{d}{dt}e^{tX} = Xe^{tX}$; in particular $\frac{d}{dt}\big|_{t=0} e^{tX} = X$.

Proof (sketch). (i) is clear. For (ii), multiply the two absolutely convergent series and use the binomial theorem, which holds for commuting matrices. For (iii), conjugate the series term by term. For (iv), by (iii) and Theorem 7.5.2 we may assume X is upper triangular with diagonal entries $\lambda_1, \dots, \lambda_n$; then e^X is upper triangular with diagonal entries $e^{\lambda_1}, \dots, e^{\lambda_n}$, so $\det e^X = e^{\lambda_1 + \cdots + \lambda_n} = e^{\operatorname{Tr} X}$. For (v), differentiate the series term by term. □

17.3.3 Lie Algebras

A Lie group is a curved object, but its tangent space at the identity is a vector space that retains much of the structure of the group. The extra structure it carries is a Lie bracket.

Definition 17.3.5 (Lie algebra)

A ***Lie algebra*** over a field k is a k-vector space $\mathfrak{g}$ with a bilinear map $[\cdot,\cdot]:\mathfrak{g}\times\mathfrak{g}\to\mathfrak{g}$, the ***Lie bracket***, such that:

(i) $[X,X]=0$ for all $X\in\mathfrak{g}$;

(ii) the ***Jacobi identity*** $[X,[Y,Z]]+[Y,[Z,X]]+[Z,[X,Y]]=0$ holds for all $X,Y,Z\in\mathfrak{g}$.

Expanding $[X+Y,X+Y]=0$ shows that (i) implies $[X,Y]=-[Y,X]$; when $\operatorname{char}k\neq 2$, the two conditions are equivalent.

NOTE On any associative algebra, such as $k^{n\times n}$ or $\operatorname{End}(V)$, the commutator $[X,Y]=XY-YX$ is a Lie bracket: $[X,X]=0$ is clear, and the Jacobi identity follows by expanding the triple brackets, all twelve terms cancelling in pairs. With this bracket, $k^{n\times n}$ is the Lie algebra $\mathfrak{gl}_n(k)$, and $\operatorname{End}(V)$ is $\mathfrak{gl}(V)$.

Definition 17.3.6 (Lie algebra of a matrix Lie group)

For a matrix Lie group $G\subseteq \mathrm{GL}_n(\mathbb{C})$, the ***Lie algebra*** of G is

$$\mathfrak{g}=\operatorname{Lie}(G)=\{X\in\mathbb{C}^{n\times n}:e^{tX}\in G\text{ for all }t\in\mathbb{R}\}.$$

We state without proof that $\mathfrak{g}$ is a real vector subspace of $\mathbb{C}^{n\times n}$ closed under the commutator bracket, so that it is a real Lie algebra, and that it is the tangent space T_IG of G at the identity. By Proposition 17.3.4(v), each $X\in\mathfrak{g}$ is the velocity at $t=0$ of the curve $t\mapsto e^{tX}$ in G.

Example 17.3.7 (Classical Lie algebras)

- $\mathfrak{gl}_n(\mathbb{R})=\mathbb{R}^{n\times n}$, of dimension n^2.
- $\mathfrak{sl}_n(\mathbb{R})=\{X\in\mathbb{R}^{n\times n}:\operatorname{Tr}X=0\}$, of dimension n^2-1: by Proposition 17.3.4(iv), $\det e^{tX}=e^{t\operatorname{Tr}X}$ equals 1 for all t exactly when $\operatorname{Tr}X=0$.
- $\mathfrak{so}(n)=\{X\in\mathbb{R}^{n\times n}:X^t=-X\}$, the skew-symmetric matrices, of dimension $\frac{n(n-1)}{2}$; differentiating $(e^{tX})^te^{tX}=I$ at $t=0$ gives $X^t+X=0$.
- $\mathfrak{su}(n)=\{X\in\mathbb{C}^{n\times n}:X^*=-X,\ \operatorname{Tr}X=0\}$, the traceless skew-Hermitian matrices, of real dimension n^2-1.
- $\mathfrak{sp}(2n,\mathbb{R})=\{X\in\mathbb{R}^{2n\times 2n}:X^tJ+JX=0\}$, of dimension $n(2n+1)$: the condition says that JX is symmetric, and $X\mapsto JX$ is a bijection.

The ***exponential map*** of G is $\exp : \mathfrak{g} \to G$, $X \mapsto e^X$.

NOTE Near the identity, the exponential map translates between G and $\mathfrak{g}$. For X, Y near 0, the Baker–Campbell–Hausdorff formula

$$e^X e^Y = e^{X+Y+\frac{1}{2}[X,Y]+\cdots}$$

holds, where the omitted terms are iterated brackets of X and Y; so multiplication in G near the identity is determined by the bracket of $\mathfrak{g}$. By the inverse function theorem, exp is a diffeomorphism from a neighborhood of $0 \in \mathfrak{g}$ onto a neighborhood of $I \in G$, since its derivative at 0 is the identity. It need not be surjective: the matrix $\left(\begin{smallmatrix} -1 & 1 \\ 0 & -1 \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{R})$ is not e^X for any $X \in \mathfrak{sl}_2(\mathbb{R})$. For compact connected groups such as $\mathrm{SO}(n)$ and $\mathrm{SU}(n)$, however, exp is surjective.

17.3.4 Homomorphisms and Representations

Definition 17.3.8 (Lie algebra homomorphism)

A linear map $\varphi : \mathfrak{g} \to \mathfrak{h}$ between Lie algebras is a ***homomorphism*** if $\varphi([X, Y]) = [\varphi(X), \varphi(Y)]$ for all $X, Y \in \mathfrak{g}$.

We state the next two facts without proof.

Proposition 17.3.9 (The differential of a homomorphism)

Let $\Phi : G \to H$ be a continuous homomorphism of matrix Lie groups. There is a unique real-linear map $\mathrm{d}\Phi : \mathfrak{g} \to \mathfrak{h}$ with $\Phi(e^{tX}) = e^{t\,\mathrm{d}\Phi(X)}$ for all $X \in \mathfrak{g}$ and $t \in \mathbb{R}$, and $\mathrm{d}\Phi$ is a Lie algebra homomorphism. Moreover $\mathrm{d}(\Psi \circ \Phi) = \mathrm{d}\Psi \circ \mathrm{d}\Phi$ and $\mathrm{d}(\mathrm{id}_G) = \mathrm{id}_{\mathfrak{g}}$, so $G \mapsto \mathfrak{g}$, $\Phi \mapsto \mathrm{d}\Phi$ is a functor (Definition 8.2.1) from matrix Lie groups to real Lie algebras.

NOTE Conversely, if G is connected and simply connected (every loop in G can be contracted to a point), every Lie algebra homomorphism $\mathfrak{g} \to \mathfrak{h}$ is $\mathrm{d}\Phi$ for a unique $\Phi : G \to H$. Together with Lie's third theorem (every finite-dimensional real Lie algebra is the Lie algebra of a Lie group), this makes the category of simply connected Lie groups equivalent to that of finite-dimensional real Lie algebras. These statements concern Lie groups in general, which need not be matrix groups.

Definition 17.3.10 (Representation of a Lie algebra)

A ***representation*** of a Lie algebra $\mathfrak{g}$ on a vector space V is a Lie algebra homomorphism $\rho : \mathfrak{g} \to \mathfrak{gl}(V)$, that is, a linear map with $\rho([X,Y]) = \rho(X)\rho(Y) - \rho(Y)\rho(X)$.

By Proposition 17.3.9, a continuous representation $G \to \mathrm{GL}(V)$ of a matrix Lie group induces a representation of $\mathfrak{g}$ on V, its differential.

Example 17.3.11 (The adjoint representation)

Every Lie algebra $\mathfrak{g}$ has the ***adjoint representation*** $\mathrm{ad} : \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g})$, $\mathrm{ad}(X)(Y) = [X,Y]$. It is a homomorphism because of the Jacobi identity: for $Z \in \mathfrak{g}$,

$$\begin{aligned}[\mathrm{ad}\,X, \mathrm{ad}\,Y](Z) &= [X,[Y,Z]] - [Y,[X,Z]] \\ &= -[Z,[X,Y]] = [[X,Y],Z] = \mathrm{ad}([X,Y])(Z),\end{aligned}$$

where the second equality is the Jacobi identity rewritten using antisymmetry.

17.3.5 Structure and Classification

The structure theory of Lie algebras parallels that of finite groups, with ideals in the role of normal subgroups. An ***ideal*** of a Lie algebra $\mathfrak{g}$ is a subspace $\mathfrak{i}$ with $[X,Y] \in \mathfrak{i}$ for all $X \in \mathfrak{g}$ and $Y \in \mathfrak{i}$. For subspaces $\mathfrak{a}, \mathfrak{b}$, let $[\mathfrak{a},\mathfrak{b}]$ be the span of the brackets $[X,Y]$ with $X \in \mathfrak{a}$, $Y \in \mathfrak{b}$. A Lie algebra is ***abelian*** if all brackets vanish.

Definition 17.3.12 (Solvable, semisimple, and simple Lie algebras)

Let $\mathfrak{g}$ be a finite-dimensional Lie algebra. Its ***derived series*** is $\mathfrak{g}^{(0)} = \mathfrak{g}$, $\mathfrak{g}^{(k+1)} = [\mathfrak{g}^{(k)}, \mathfrak{g}^{(k)}]$.

- $\mathfrak{g}$ is ***solvable*** if $\mathfrak{g}^{(k)} = 0$ for some k.
- $\mathfrak{g}$ is ***semisimple*** if its only solvable ideal is 0.
- $\mathfrak{g}$ is ***simple*** if it is not abelian and its only ideals are 0 and $\mathfrak{g}$.

NOTE We state the main structure theorems without proof. The sum of two solvable ideals is a solvable ideal, so $\mathfrak{g}$ has a largest solvable ideal $\mathfrak{r}$, its ***radical***, and $\mathfrak{g}$ is semisimple exactly when $\mathfrak{r} = 0$. Over a field of characteristic 0, $\mathfrak{g} = \mathfrak{r} \oplus \mathfrak{s}$ as vector spaces for a semisimple subalgebra $\mathfrak{s}$ (the

Levi decomposition), and every semisimple Lie algebra is, uniquely, a direct sum of simple ideals.
The simple Lie algebras over $\mathbb{C}$ are classified by Dynkin diagrams:

- four infinite families, $A_n = \mathfrak{sl}_{n+1}(\mathbb{C})$ $(n \geqslant 1)$, $B_n = \mathfrak{so}_{2n+1}(\mathbb{C})$ $(n \geqslant 2)$, $C_n = \mathfrak{sp}_{2n}(\mathbb{C})$ $(n \geqslant 3)$, and $D_n = \mathfrak{so}_{2n}(\mathbb{C})$ $(n \geqslant 4)$;
- five exceptional algebras, G_2, F_4, E_6, E_7, E_8, of dimensions 14, 52, 78, 133, 248.

The classification proceeds through root systems. Given a semisimple $\mathfrak{g}$ and a Cartan subalgebra $\mathfrak{h}$ (a maximal abelian subalgebra consisting of semisimple elements), the adjoint action of $\mathfrak{h}$ decomposes $\mathfrak{g}$ into root spaces; the resulting set of roots $\Phi \subset \mathfrak{h}^*$ is a root system, which is encoded by its Dynkin diagram. This is one of the great classification theorems of mathematics (Killing, 1888; Cartan, 1894).

Example 17.3.13 ($\mathfrak{sl}_2(\mathbb{C})$, the simplest simple Lie algebra)

The Lie algebra $\mathfrak{sl}_2(\mathbb{C})$ of traceless 2×2 complex matrices has dimension 3, with standard basis

$$e = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad f = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \qquad h = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

satisfying $[h, e] = 2e$, $[h, f] = -2f$, and $[e, f] = h$ (Exercise 17.3.14(1)). It is the Lie algebra of type A_1. Its finite-dimensional representations are completely known: for each $n \geqslant 0$ there is, up to isomorphism, exactly one irreducible representation of dimension $n + 1$, realized on $\mathrm{Sym}^n(\mathbb{C}^2)$ (Definition 9.10.3). The representation theory of $\mathfrak{sl}_2$ is the foundation of the general theory.

Exercise 17.3.14 (Exercises)

(1) Verify the relations $[h, e] = 2e$, $[h, f] = -2f$, $[e, f] = h$ for the standard basis of $\mathfrak{sl}_2(\mathbb{C})$.
(2) Show that $\mathfrak{so}(3)$, the real skew-symmetric 3×3 matrices with the commutator bracket, is isomorphic as a Lie algebra to $\mathbb{R}^3$ with the cross product as bracket.
(3) The ***Killing form*** of a finite-dimensional Lie algebra $\mathfrak{g}$ is $B(X, Y) = \mathrm{Tr}(\mathrm{ad}\, X \circ \mathrm{ad}\, Y)$. Show that for $\mathfrak{g} = \mathfrak{gl}_n(\mathbb{C})$,

$$B(X, Y) = 2n\,\mathrm{Tr}(XY) - 2\,\mathrm{Tr}\,X\,\mathrm{Tr}\,Y.$$

Solution. **(1)** Directly:

$$[h,e] = he - eh = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right) - \left(\begin{smallmatrix} 0 & -1 \\ 0 & 0 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 0 & 2 \\ 0 & 0 \end{smallmatrix}\right) = 2e,$$
$$[h,f] = hf - fh = \left(\begin{smallmatrix} 0 & 0 \\ -1 & 0 \end{smallmatrix}\right) - \left(\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right) = -2f,$$
$$[e,f] = ef - fe = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right) - \left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right) = h.$$

(2) A basis of $\mathfrak{so}(3)$ is

$$E_1 = \left(\begin{smallmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{smallmatrix}\right), \qquad E_2 = \left(\begin{smallmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{smallmatrix}\right), \qquad E_3 = \left(\begin{smallmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{smallmatrix}\right),$$

and a direct computation gives $[E_1, E_2] = E_3$, $[E_2, E_3] = E_1$, $[E_3, E_1] = E_2$. The standard basis of $\mathbb{R}^3$ satisfies $e_1 \times e_2 = e_3$, $e_2 \times e_3 = e_1$, $e_3 \times e_1 = e_2$. Let $\varphi : \mathfrak{so}(3) \to \mathbb{R}^3$ be the linear isomorphism with $\varphi(E_i) = e_i$. Both brackets are bilinear and antisymmetric, so $\varphi([X,Y]) = \varphi(X) \times \varphi(Y)$ holds for all X, Y once it holds for pairs of basis elements, which is what we checked. So φ is an isomorphism of Lie algebras.

(3) For $A, C \in \mathfrak{gl}_n(\mathbb{C})$, the linear map $Z \mapsto AZC$ on $\mathfrak{gl}_n(\mathbb{C})$ has trace $\operatorname{Tr} A \cdot \operatorname{Tr} C$: on the basis of matrix units E_{ij}, the coefficient of E_{ij} in $AE_{ij}C$ is $A_{ii}C_{jj}$, and $\sum_{i,j} A_{ii}C_{jj} = \operatorname{Tr} A \operatorname{Tr} C$. Now

$$\operatorname{ad} X \circ \operatorname{ad} Y(Z) = X(YZ - ZY) - (YZ - ZY)X = XYZ - XZY - YZX + ZYX.$$

The four terms have the form AZC with $(A, C) = (XY, I)$, (X, Y), (Y, X), and (I, YX), so

$$B(X,Y) = n \operatorname{Tr}(XY) - \operatorname{Tr} X \operatorname{Tr} Y - \operatorname{Tr} Y \operatorname{Tr} X + n \operatorname{Tr}(YX) = 2n \operatorname{Tr}(XY) - 2 \operatorname{Tr} X \operatorname{Tr} Y.$$

For $X, Y \in \mathfrak{sl}_n(\mathbb{C})$, the map $\operatorname{ad} X \circ \operatorname{ad} Y$ preserves $\mathfrak{sl}_n(\mathbb{C})$ and kills I, and $\mathfrak{gl}_n(\mathbb{C}) = \mathfrak{sl}_n(\mathbb{C}) \oplus \mathbb{C}I$; so the Killing form of $\mathfrak{sl}_n(\mathbb{C})$ is $B(X,Y) = 2n \operatorname{Tr}(XY)$. It is nondegenerate: if $X \neq 0$ is traceless, so is X^*, and $\operatorname{Tr}(XX^*) = \sum_{i,j} |X_{ij}|^2 > 0$. By Cartan's criterion, which we do not prove (a finite-dimensional Lie algebra over $\mathbb{C}$ is semisimple if and only if its Killing form is nondegenerate), $\mathfrak{sl}_n(\mathbb{C})$ is semisimple.