

Vol. 5 – MATHEMATICAL METHODS

The Undergraduate Companion to Theoretical Physics

Andrea Kouta Dagnino[‡]

[‡]*Open University, Milton Keynes, UK.*

E-mail: k.y.dagnino@gmail.com

Contents

I Analysis	4		
1 Unit A1: Sets, functions and vectors	5	7.2 Telescoping series	50
2 Unit A2: Number systems	6	7.3 Manipulating series	51
3 Unit A3: Mathematical language and proofs (why haven't you finished this yet!!)	7	7.4 Non-negative series	53
3.1 Mathematical statements . . .	7	7.5 Series with positive and negative terms	60
3.2 Direct Proof	10		
4 Unit A4: Real functions, graphs and conics(why haven't you finished this yet!!)	16	8 Unit D4: Functions and Continuity	64
5 Unit D1: Numbers	17	8.1 Real functions	64
5.1 The set \mathbb{N} of Natural Numbers	17	8.2 Continuity	65
5.2 The Set \mathbb{Q} of Rational Numbers	20	8.3 Properties of continuous functions	67
5.3 The Set \mathbb{R} of Real Numbers .	21	8.4 Trigonometric and exponential functions	70
5.4 Absolute Value	24		
5.5 The Completeness Axiom .	25	9 Unit F1: Limits	73
6 Unit D2: Sequences	28	9.1 Introduction to limits of functions	73
6.1 Introduction to sequences . .	28	9.2 Asymptotic behaviour	78
6.2 Convergence of sequences . .	29	9.3 Continuity of functions	81
6.3 Formal Proofs of Limit Theorems	31	9.4 Unusual function continuity .	83
6.4 Null sequences	33	9.5 Uniform continuity	85
6.5 Limit theorems for Convergent sequences	37	10 Unit F2: Differentiation	88
6.6 Divergent sequences	43	10.1 Continuity and differentiability	92
6.7 Monotone convergence theorem	47	10.2 Rules of differentiation	92
7 Series	49	10.3 Rolle's theorem and local extrema	96
7.1 Introduction to Series	49	10.4 Mean value theorem	97
		10.5 L'Hopital's rule	100
		11 Unit F3: Integration	102
		11.1 The Riemann integral	102
		11.2 Inequalities and series with integrals	108
		12 Unit F4: Power series	115

12.1 Taylor series	115	17.4 Normal subgroups	205
12.2 Convergence	121	18 Unit E2: Quotient groups and conjugacy	208
12.3 The combination rules	125	18.1 Quotient groups	208
II Algebra and Group Theory	131	18.2 Quotient group of infinite groups	211
13 Unit B1: Symmetry and groups	132	18.3 Conjugacy	213
13.1 Symmetry in \mathbb{R}^2	132	18.4 Normal subgroups and conjugacy	216
13.2 Representing symmetries	136	18.5 Conjugacy in $S(\mathcal{F})$	221
13.3 Definition of a Group	137	19 Unit E3: Homomorphisms	228
13.4 Properties of groups and group elements	142	19.1 Image and kernels	233
13.5 Symmetry in \mathbb{R}^3	143	19.2 First isomorphism theorem	236
13.6 The Dihedral group	147	20 Unit E4: Group actions	242
14 Unit B2: Subgroups and isomorphisms	150	20.1 What are group actions?	242
14.1 Subgroups	150	20.2 Orbits and stabilisers	246
14.2 Cyclic groups and subgroups	156	20.3 The Orbit-Stabiliser theorem	251
14.3 Cyclic groups and modular arithmetic	159	20.4 The Counting theorem	256
14.4 Isomorphisms	161	21 Sylow theorems	261
14.5 Standard groups	165	22 Rings	262
14.6 Direct product of groups	166	23 Polynomials	263
15 Unit B3: Permutations	169	24 Modules	264
15.1 Permutations	169	III Representation Theory and Lie Algebra	265
15.2 Permutation groups	173	IV Differential Equations (get the PDE chapters working)	266
15.3 Even and Odd symmetries	175	25 Fundamentals	267
15.4 Conjugacy of S_n	180	25.1 Definitions	267
15.5 Subgroups of S_4	182	25.2 Integral formulation	268
15.6 Cayley's Theorem	184	25.3 Picard iteration	270
16 Unit B4: Lagrange's Theorem and small groups	187	25.4 Existence and uniqueness	273
16.1 Lagrange's Theorem	187	26 First Order ODEs	274
16.2 Groups of small order	189	26.1 Exact Differential Equations	274
17 Unit E1: Cosets and normal subgroups	194		
17.1 Matrix groups	194		
17.2 Cosets	199		
17.3 Right cosets	202		

26.2 Separable Differential Equations	275	32 Sturm-Liouville theory	306
26.3 Inexact Differential Equations	275	33 Distributions	307
26.4 Integrating Factor Method .	277	33.1 Introducing the Dirac delta .	307
26.5 Bernoulli Equations	277	33.2 Rigorous treatment	309
26.6 Stability and Equilibrium points	278	33.3 Distributions	310
27 Second Order ODEs	279	34 Laplace transform methods	311
27.1 Homogeneous equation	279	34.1 Basic definition and properties of the Laplace transform	311
27.2 Non-homogeneous	282	34.2 Solving ODEs with Laplace transforms	314
27.3 Undetermined Coefficients .	283		
27.4 Variation of Constants	283		
27.5 Reduction of Order	284		
28 Mechanical Vibrations and Resonance Phenomena	285	35 Phase plane analysis	315
28.1 Homogeneous Equation	285		
28.2 Damped Harmonic Motion .	286		
28.3 Forced Oscillations	288	36 First order PDEs	316
28.4 Resonance	290		
29 General Linear ODEs	291	37 Second order PDEs: a general overview	317
29.1 Existence and Uniqueness .	291		
29.2 Fundamental set and Wronskians	291	38 Parabolic PDEs: Diffusion	318
29.3 Homogeneous ODE	292		
29.4 Non-homogeneous equation .	293	39 Hyperbolic PDEs: Waves	319
29.5 Higher Order ODEs	294		
30 Systems of Linear Differential Equations	296	40 Elliptic PDEs: Electrostatics	320
30.1 Non-degenerate Eigenvalues .	296		
30.2 Matrix exponentiation	298	V Linear Algebra	321
30.3 Higher Order Constant Coefficient Equations	298		
30.4 Triangulation	299	41 Vector spaces	322
30.5 Jordan Form	300	41.1 Definitions	322
31 Series Solutions and special functions	302	41.2 Basis and dimensions	324
31.1 Power Series	302	41.3 Operations on subspaces . . .	327
31.2 Series Solutions near ordinary points	303	42 Euclidean geometry in \mathbb{R}^3	334
31.3 Airy's Equation	304		
		43 Matrix algebra	335
		44 Linear transformations	336
		44.1 What is a map?	336
		44.2 What is a linear map?	340
		44.3 Isomorphisms	344
		44.4 Linear maps and matrices . .	344
		44.5 Change of basis and equivalence	351
		45 Solving linear equations	356

45.1 Structure of solutions	356	49.7 Application to special relativity: four-vectors	417
45.2 Elementary matrix operations	358		
45.3 Inverting matrices	361		
46 Determinants	366	50 Tensor calculus	419
46.1 The determinant of a matrix	366	50.1 Christoffel symbols	419
46.2 Laplace expansion	370	50.2 Differentiating tensors	420
46.3 Cramer's rule	374	50.3 Application to geometry: curvilinear coordinates	422
47 Inner product spaces	376	50.4 Geodesics	423
47.1 Inner products	376		
47.2 Projectors	377		
47.3 Inner products and matrices .	378		
47.4 Bilinear and Sesquilinear forms	379		
48 Eigen-everything	383	VI Differential Geometry	424
48.1 Finding eigenvalues and eigenvectors	383	51 <i>Differentiable Manifolds</i>	425
48.2 Matrix diagonalization	386	52 <i>Differential forms</i>	426
48.3 Orthogonal diagonalization .	388	53 <i>Integrating on manifolds</i>	427
48.4 Classifying conics	393	54 <i>Curvature</i>	428
48.5 Matrix exponentials and Lie algebras	396	55 <i>Lie derivatives</i>	429
48.6 Schur's triangulation theorem	397		
48.7 Jordan canonical form (make sure to write by October) .	401		
49 Tensor algebra	402	VII Complex analysis	430
49.1 Einstein summation convention	402		
49.2 Cartesian tensors	403	VIII Calculus of Variations	431
49.3 The δ_{ij} and ϵ_{ijk} tensors . .	410		
49.4 Physical examples of carte- sian tensors	410	IX Fourier Analysis	432
49.5 Non-cartesian tensors	411		
49.6 Covariance and contravariance	414	X Functional Analysis and Operator theory	433
		<i>Bibliography</i>	435

*God used beautiful mathematics in creating
the world.*

— P.A.M. Dirac



References

Several textbooks, online courses/resources were referenced heavily (to the extend of making this text completely unoriginal, yet helpful for revision) throughout the writing of these lecture notes. We list the most relevant below:

- The Open University textbooks
- Lang, "*Undergraduate analysis*"
- Ross, "*Elementary analysis*"
- Armstrong, "*Groups and symmetries*"
- Stewart and Tall, "*Complex analysis*"
- Riley, Hobson and Bence, "*Mathematical methods for physics and engineering*"
- Lang, "*Linear algebra*"
- Kammler, "*A first course in Fourier analysis*"
- Fulton and Harris, "*Representation theory, a first course*"
- Jevanjee, "*An introduction to Tensors and Group Theory for Physicists*"
- Collins, "*Differential and Integral equations*"

Part I

Analysis

Unit A1: Sets, functions and vectors

- (i) $x + y = y + x, \forall x, y \in \mathbb{R}$
- (ii) $x \cdot y = y \cdot x, \forall x, y \in \mathbb{R}$

Unit A2: Number systems

2

Unit A3: Mathematical language and proofs (why haven't you finished this yet!!)

3.1 Mathematical statements

Types of statements

Assertions are a fundamental concept, a sort of axiom that we will not define rigorously, but can be seen as any form of some statement. For example, $\{1, 2\}$ is greater than 2 is an assertion.

A **proposition** P is a statement that can be either true or false. $\{1, 2\}$ is greater than 2 is not a proposition, since it is neither true nor false, it just doesn't make sense. A statement $P(x)$ which is either true or false depending on the value of some variable(s) then it is a **variable proposition**.

A **theorem** is a mathematical statement that is (in general not obviously) true. Less important theorems are also called **propositions** (note that propositions can have two meanings depending on the context they're used in). A **lemma** is a small theorem, a result that can be used to prove other theorems. A **corollary** is a theorem that follows from another theorem.

Logical connectives

All statements P have a related statement called **negation** $\neg P$, whose truth table is: For

P	$\neg P$
T	F
F	T

example for the following variable proposition:

$$P(x) : x \leq 0 \quad (3.1.1)$$

the negation is:

$$\neg P(x) : x > 0 \quad (3.1.2)$$

Similarly, for the proposition

$$Q : \text{there are at least 10 two-digit numbers less than 20} \quad (3.1.3)$$

$$\neg Q : \text{there are at most 9 two-digit numbers less than 20} \quad (3.1.4)$$

Given two statements P, Q we can insert the word *and*, *or* in between to give a new statement, the **conjunction** $P \wedge Q$ (P and Q) and **disjunction** $P \vee Q$ (P or Q) respectively. Their truth tables are: So \wedge is false whenever at least one of the two statements is false,

P	Q	$P \wedge Q$	$P \vee Q$
T	T	T	T
T	F	F	T
F	T	F	T
F	F	F	F

whereas \vee is false only when both statements are false. For example given $P : 2 \text{ is prime}$ and $Q : 2 \text{ is even}$ then:

$$P \wedge Q : 2 \text{ is prime and } 2 \text{ is even} \quad (3.1.5)$$

$$P \vee Q : 2 \text{ is prime or } 2 \text{ is even} \quad (3.1.6)$$

We can also negate conjunctions and disjunctions. For example, given the conjunction

$$p \text{ is odd } \underline{\text{and}} \text{ prime} \quad (3.1.7)$$

its negation is

$$p \text{ is even } \underline{\text{or}} \text{ not prime.} \quad (3.1.8)$$

Moreover, given the disjunction

$$\underline{\text{Either}} \ A = B \ \underline{\text{or}} \ A \cup B = \emptyset \quad (3.1.9)$$

its negation is

$$\underline{\text{Both}} \ A \neq B \ \underline{\text{and}} \ A \cup B \neq \emptyset. \quad (3.1.10)$$

It is therefore clear that when performing a negation *and* \longleftrightarrow *or* and *both* \longleftrightarrow *either*. Therefore:

$$\neg(P \wedge Q) = \neg P \vee \neg Q \quad \neg(P \vee Q) = \neg P \wedge \neg Q \quad (3.1.11)$$

Implications

An **implication** $P \implies Q$ (or P is sufficient for Q , P only if Q) is a mathematical statement of causality between two propositions P, Q in the form *if P then Q*. Its truth table is therefore:

P	Q	$P \implies Q$
T	T	T
T	F	F
F	T	T
F	F	T

so it is only false when the conclusion is false despite the hypothesis being true. For example:

$$x > 2 \text{ only if } x > 4 \quad (3.1.12)$$

can be expressed as $x > 2 \implies x > 4$. To negate an implication, we use a conjunction. Indeed, an implication states *if P then Q*, so its opposite would be to state that it is the case that P is true and Q is false, in other words P and not Q .

For example consider the implication:

$$(m \text{ divides } 12) \implies (m \text{ divides } 3 \text{ or } m \text{ divides } 4) \quad (3.1.13)$$

then its negation is:

$$(m \text{ divides } 12) \wedge (m \text{ does not divide neither } 3 \text{ nor } 4) \quad (3.1.14)$$

To every implication $P \implies Q$ is an associated **converse** which states $Q \implies P$. It is important to remember that whether or not an implication is true tells you nothing about the truth of its converse. Indeed, consider the implication:

$$\text{If } m, n \text{ are both odd, then } m + n \text{ is even} \quad (3.1.15)$$

which is clearly true, then its converse:

$$\text{If } m + n \text{ is even, then } m, n \text{ are both odd} \quad (3.1.16)$$

which is not necessarily true.

To every implication $P \implies Q$ is an associated *contrapositive* $\neg Q \implies \neg P$ to which it is equivalent (therefore they state the same thing). For example the contrapositive of the previous implication 3.1.14 is:

$$\text{If } m + n \text{ is odd, then } m, n \text{ are not both odd} \quad (3.1.17)$$

Equivalences

The statement $P \implies Q \wedge Q \implies P$ is stated mathematically as:

$$P \iff Q \quad (3.1.18)$$

and is called an **equivalence**, and can be expressed as P if and only if (iff) Q . One common exception is in definitions, where simply *if* is used instead of *iff*. Equivalences can also be thought of as a conjunction between $P \implies Q$ and $\neg P \implies \neg Q$. So for example:

$$m \text{ is even} \text{ iff } m^2 \text{ is even} \quad (3.1.19)$$

is the same as stating

$$(m \text{ is even} \implies m^2 \text{ is even}) \wedge (m \text{ is odd} \implies m^2 \text{ is odd}) \quad (3.1.20)$$

Universal and existential quantifiers

The **universal quantifier** \forall can be expressed as *for all/every*. So, the statement:

$$\forall x \in \mathbb{R}, P(x) \quad (3.1.21)$$

expresses the proposition $P(x)$ for all $x \in \mathbb{R}$. Mathematical statements including the words *there are no, for all/every/each, every, any* are usually **universal statements** because they state the validity of a proposition for a certain set of values for a variable.

The **existential quantifier** \exists can be expressed as *there exists*. So the statement:

$$\exists x \in \mathbb{R}, P(x) \quad (3.1.22)$$

expresses the proposition $P(x)$ is true for some $x \in \mathbb{R}$.

It is then clear that the negation of an existential statement is a universal statement and vice versa. Indeed the negation of $\forall x \in \mathbb{R}, P(x)$ is:

$$\exists x \in \mathbb{R}, \neg P(x) \quad (3.1.23)$$

and the negation of $\exists x \in \mathbb{R}, P(x)$ is:

$$\forall x \in \mathbb{R}, \neg P(x) \quad (3.1.24)$$

3.2 Direct Proof

Proof by exhaustion

If n is an odd number between 0 and 10, then n^2 is also odd.

Proof. The odd numbers between 0 and 10 are 1,3,5,7,9, and their respective squares are 1,9,25,49,81 which are all odd. ■

This is a **proof by exhaustion**, which consists in proving the statement for all the possible values of the variable in question, in this case n .

Another method of proof is simply algebraic verification.

[Geometric series identity] Let $a, b \in \mathbb{R}$ and let n be a positive integer. Then:

$$a^n - b^n = (a - b)(a^{n-1} + a^{n-2}b + \dots + ab^{n-2} + b^{n-1}) \quad (3.2.1)$$

Proof. Expanding the RHS gives:

$$a^n + a(a^{n-2}b + a^{n-3}b^2 + \dots + ab^{n-2}) - b(a^{n-2}b + a^{n-3}b^2 + \dots + ab^{n-2}) - b^n \quad (3.2.2)$$

$$= a^n - b^n + b(a^{n-1} + a^{n-2}b + \dots + a^2b^{n-1}) - b(a^{n-2}b + a^{n-3}b^2 + \dots + ab^{n-2}) \quad (3.2.3)$$

$$= a^n - b^n \quad (3.2.4)$$

■

Proving implications

To prove an implication $P \implies Q$, often we assume that P is true. We then find some true statement

$$P \implies P_1 \quad (3.2.5)$$

from which we deduce that P_1 is true. We then find some other true statement:

$$P_1 \implies P_2 \quad (3.2.6)$$

from which we deduce that P_2 is true. We continue this process, until we find that P_n is true and

$$P_n \implies Q \quad (3.2.7)$$

which proves that Q is true, whenever P is true.

Prove that if n is odd, then n^2 is odd.

Proof. Let n be an odd integer, then:

$$\exists k \in \mathbb{Z} \text{ s.t. } n = 2k + 1 \implies n^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1 \quad (3.2.8)$$

which is in the form $2m + 1$ with $m = 2k^2 + 2k$. Therefore n^2 is odd whenever n is odd. ■

We now state an important theorem that will help us in proving other propositions.

[Fundamental Theorem of Arithmetic] Every integer greater than 1 can be written as a unique product of prime numbers.

$$\forall n \in \mathbb{Z}, n^3 + 3n^2 + 2n \text{ is divisible by 6.}$$

Proof. Let $n \in \mathbb{Z}$ then:

$$n^3 + 3n^2 + 2n = n(n^2 + 3n + 2) = n(n+1)(n+2) \quad (3.2.9)$$

hence $n^3 + 3n^2 + 2n$ is the product of three consecutive integers, which implies that one of them must be divisible by 2, and another must be divisible by 3. Therefore both 2 and 3 divide $n^3 + 3n^2 + 2n$. Hence by the Fundamental Theorem of Arithmetic we may write:

$$n^3 + 3n^2 + 2n = 2 \cdot 3 \cdot r = 6r \quad (3.2.10)$$

for some r . We may then deduce that 6 divides $n^3 + 3n^2 + 2n$. ■

Consider the following proof:

$$1 = -1 \implies 1^2 = (-1)^2 \implies 1 = 1 \quad (3.2.11)$$

therefore, since it is true that $1 = 1$, we can deduce that $1 = -1$.

This is clearly wrong, because we started with the statement we wanted to prove, and through a series of implications, we found a true statement. However, this does not achieve anything, since implications only work one way, we have proved that if $1 = -1$ then $1 = 1$, but not the converse.

More generally, if $P \implies Q$, then we have no information about Q when P is false, as can be seen from the truth table.

Proving equivalences

Recall that an equivalence $P \iff Q$ is the same as saying $(P \implies Q) \wedge (Q \implies P)$. Thus, if we wish to prove $P \iff Q$ then generally it suffices to prove both $P \implies Q$ and $Q \implies P$. When dealing with equations or inequalities, often the individual implications $P_i \implies P_{i+1}$ are reversible, so they are actually equivalences. In that case we can simply prove $P \iff Q$ using a series of equivalences:

$$P \iff P_1 \iff P_2 \iff \dots \iff P_n \iff Q \quad (3.2.12)$$

$$x(x-2) = 3 \iff x = -1 \vee x = 3$$

Proof. We will use a series of equivalences:

$$x(x - 2) = 3 \iff x^2 - 2x - 3 = 0 \quad (3.2.13)$$

$$\iff (x + 1)(x - 3) = 0 \quad (3.2.14)$$

$$\iff (x + 1 = 0) \vee (x - 3 = 0) \quad (3.2.15)$$

$$\iff x = -1 \vee x = 3 \quad (3.2.16)$$

as required. ■

For two arbitrary sets A, B , $A \cup B = A \iff B \subseteq A$

Proof. We will use the approach of tackling the implication and its converse separately.

\implies We firstly prove the rightward implication. Assume that $A \cup B = A$, and let $x \in B$, then it follows that $x \in A \cup B$, and thus $x \in A$. So, $B \subseteq A$, since $x \in B \implies x \in A$.

\impliedby we prove the leftward implication. Assume that $B \subseteq A$.

Then let $x \in A \cup B$ so that $x \in A \vee x \in B$. If $x \in B$, then $x \in A$ by assumption. We can then deduce that $x \in A$, so $A \cup B \subseteq A$. Clearly we must also have $A \subseteq A \cup B$, so that $A = A \cup B$ as required.

Finally, the equivalence has been proven:

$$A \cup B = A \iff B \subseteq A \quad (3.2.17) \quad \text{■}$$

Now consider the following (incorrect) proof:

$$n \text{ is a multiple of } 5 \iff \exists k \in \mathbb{Z}, \text{s.t. } n = 5k \quad (3.2.18)$$

$$\iff \exists k \in \mathbb{Z}, \text{s.t. } n^2 = 25k^2 \quad (3.2.19)$$

$$\iff \exists k \in \mathbb{Z}, \text{s.t. } n^2 = 5(5k^2) \quad (3.2.20)$$

$$\iff \exists k \in \mathbb{Z}, \text{s.t. } n^2 \text{ is a multiple of } 5. \quad (3.2.21)$$

The problem with this proof lies in the final equivalence. Indeed, the implication is true, but its converse is not immediate and requires further justification. Indeed: I

$$n^2 \text{ is a multiple of } 5 \implies \exists k \in \mathbb{Z} \ n^2 = 5l \implies l = 5k^2 \implies n^2 = 5(5k^2) \quad (3.2.22)$$

for some k , since otherwise taking the square root of $n^2 = 5l$ we would not be able to factor out the 5, and so $n \notin \mathbb{Z}$.

[Factor Theorem in \mathbb{R}] Let $p(x)$ be a real polynomial, and let $\alpha \in \mathbb{R}$. Then $p(\alpha) = 0 \iff (x - \alpha)$ is a factor of $p(x)$.

Proof. Throughout the proof we assume that $p(x)$ is real with $\alpha \in \mathbb{R}$.

\implies Let us first prove the rightward implication. Assume that $p(\alpha) = 0$, and let:

$$p(x) = \sum_{i=0}^n a_i x^i, \quad a_n \neq 0 \quad (3.2.23)$$

Then since $p(\alpha) = 0$:

$$p(\alpha) = \sum_{i=0}^n a_i \alpha^i = 0 \implies p(x) = p(x) - p(\alpha) = \sum_{i=1}^n a_i (x^i - \alpha^i) \quad (3.2.24)$$

where the constant terms cancel out.

By the Geometric series identity, we know that $x - \alpha$ divides the above expression. Hence we have that $x - \alpha$ divides $p(x)$ as required.

\iff Now assume that $(x - \alpha)$ is a factor of $p(x)$ so that $p(x) = (x - \alpha)q(x)$. Then clearly:

$$p(\alpha) = 0 \cdot q(\alpha) = 0 \quad (3.2.25)$$

as required. ■

Proving existential and universal statements

The simplest way to prove an existential statement is to provide an object that satisfies the statement.

There exists a real positive number x such that $x \leq \sqrt{x}$.

Proof. Let $x = \frac{1}{9}$, then $x = \frac{1}{9} \leq \frac{1}{3} = \sqrt{x}$ as required. ■

We may also need to prove that a statement is false. For a universal statement, it suffices to find a **counterexample**. To prove that $P(x) \implies Q(x)$ is false we need to prove that $\exists x$ such that $P(x)$ and not $Q(x)$, this value of x is a counterexample.

Proof by induction

To prove that a mathematical statement $P(n)$ is true for $n = 1, 2, \dots$:

- (i) prove that $P(1)$ is true,

(ii) prove that $P(k) \implies P(k+1)$ for $k = 1, 2, \dots$

For all $n \geq 7$, $3^n < n!$.

Proof. (i) $P(7)$ is true since it is true that $3^7 = 2187 < 5040 = 7!$.

(ii) Assume that $P(k)$ is true for some $k \geq 7$ so that:

$$3^k < k! \quad (3.2.26)$$

Then we wish to deduce that $P(k+1)$ is true as follows:

$$3^{k+1} = 3^k \cdot 3 < 3 \cdot k! < k \cdot k! = (k+1)! \quad (3.2.27)$$

so that $3^{k+1} < (k+1)!$ as required. By mathematical induction, we conclude that $P(n)$ is true for all $n \geq 7$. ■

For $n \in \mathbb{N}$, $2^{3n+1} + 5$ is a multiple of 7.

Proof. $P(1)$ is true since $2^{3 \cdot 1 + 1} + 5 = 21 = 3 * 7$.

Assume that $P(k)$ is true for some $k \geq 1$ so that:

$$2^{3k+1} + 5 \text{ is a multiple of } 5, \quad (3.2.28)$$

Then we find that:

$$2^{3k+4} + 5 = 2^3 \cdot 2^{3k+1} + 5 = 7 \cdot 2^{3k+1} + 2^{3k+1} + 5 \quad (3.2.29)$$

which is divisible by 7 given that $P(k)$ is true. Therefore we have shown that:

$$P(k) \implies P(k+1) \quad (3.2.30)$$

for $k \in \mathbb{N}$, and by mathematical induction it follows that the proposition is true. ■

Let us now try to prove a more advanced theorem, the factor theorem presented in the previous chapter.

Let $p(x) = \sum_{i=0}^n a_i x^i$

Unit A4: Real functions, graphs and conics(why haven't you finished this yet!!)

Unit D1: Numbers

5.1 The set \mathbb{N} of Natural Numbers

Peano Axioms

We denote by \mathbb{N} the inductive set $\{1, 2, 3, \dots\}$ of all positive integers, so that each positive integer n has a successor $n + 1 = \text{succ}(n)$. We can then state the following *Peano axioms*:

N1. $1 \in \mathbb{N}$

N2. $n \in \mathbb{N} \implies \text{succ}(n) \in \mathbb{N}$

N3. $\forall n \in \mathbb{N}, 1 \neq \text{succ}(n)$

N4. $\text{succ}(n) = \text{succ}(m) \iff n = m$

N5. Let $A \subset \mathbb{N}$ which contains 1 and contains $\text{succ}(n)$ whenever it contains n , then $A = \mathbb{N}$

Remark. Assume N5 is false, then \mathbb{N} contains a set A such that:

(i) $1 \in A$

(ii) $n \in A \implies (n + 1) \in A$

(iii) $A \neq \mathbb{N}$

and consider $n_0 = \min S$, where $S = \{n \in \mathbb{N} \mid n \notin A\}$. Clearly, $n_0 \neq 1$, so n_0 is the successor of some number $n_0 - 1$. Since $n_0 \in S$, it follows that $(n_0 - 1) \in A$. However, by (ii) $(n_0 - 1) \in A \implies n_0 \in A$ which is a contradiction.

Principle of Mathematical Induction

Let P_1, P_2, \dots be a list of propositions, then the principle of mathematical induction asserts that they are true provided:

I₁ P_1 is true (basis of induction)

I₂ $(P_n \text{ is true}) \implies (P_{n+1} \text{ is true})$ (inductive step)

Proposition 5.1 (Sum of natural numbers)

The sum of the first n natural numbers is:

$$\sum_{i=1}^n = \frac{n(n+1)}{2}. \quad (5.1.1)$$

Proof. We define the n th proposition to be:

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

I₁ For the basis for induction, P_1 asserts that the sum of the first natural number is $\frac{1 \cdot 2}{2} = 1$ which is clearly true.

I₂ For the inductive step, suppose P_n is true, so we assume:

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

is true. Now:

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

so P_{n+1} is true as required. ■

Example. Let us prove that all numbers of the form $5^n - 4n - 1$ are divisible by 16, $\forall n \in \mathbb{N}$.

Proof. So the n th proposition is:

$$P_n : 5^n - 4n - 1 \text{ are divisible by 16.}$$

I₁ The basis for induction is true, since $5^1 - 4 - 1 = 0$ which is divisible by 16.

I₂ For the inductive step, suppose P_n is true, we wish to verify P_{n+1} . To do so, we write:

$$5^{n+1} - 4(n+1) - 1 = 5(5^n - 4n - 1) + 16n = 5 \cdot 16m + 16n = 16(5m + n)$$

where $5^n - 4n - 1 = 16m$, required. ■

Example. Let us prove that $|\sin nx| \leq n|\sin x|, \forall n \in \mathbb{N}, \forall x \in \mathbb{R}$.

Proof. Our n th proposition is:

$$P_n : |\sin nx| \leq n|\sin x|, \forall x \in \mathbb{R}$$

I₁ The basis for induction is clearly true, since $|\sin x| \leq |\sin x|$.

I₂ For the inductive step, assume that P_n is true, then:

$$\begin{aligned} |\sin(n+1)x| &= |\sin(nx+x)| \\ &= |\sin nx \cos x + \sin x \cos nx| \\ &\leq |\sin nx||\cos x| + |\sin x||\cos nx| \\ &\leq |\sin nx| + |\sin x| \\ &\leq n|\sin x| + |\sin x| \\ &\leq (n+1)|\sin x| \end{aligned}$$

as required, P_{n+1} holds. ■

Theorem 5.2 (Bernoulli inequality)

Let $x \in \mathbb{R}$ and $n \in \mathbb{N}$ then:

$$(1+x)^n \geq 1 + nx, \text{ when } x \geq -1 \quad (5.1.2)$$

Proof. Let $x < \geq -1$ and define:

$$P(n) : (1+x)^n \geq 1 + nx \quad (5.1.3)$$

I₁ $P(1)$ is obviously true, since the LHS reads $(1+x)^1 = (1+x)$ which is equal to the RHS.

I₂ Now let $P(k)$ be true for some $k \geq 1$, so that:

$$(1+x)^k \geq 1 + kx \quad (5.1.4)$$

It follows that:

$$(1+x)^{k+1} \geq (1+kx)(1+x) \quad (5.1.5)$$

$$\geq 1 + (k+1)x + kx^2 \quad (5.1.6)$$

$$\geq 1 + (k+1)x \quad (5.1.7)$$

since $kx^2 \geq 0$. It follows that $P(k+1)$ is true, whenever $P(k)$ is verified.

Hence, by the principle of mathematical induction, we have that

$$(1+x)^n \geq 1 + nx, \text{ when } x \geq -1 \quad (5.1.8)$$

as desired. ■

5.2 The Set \mathbb{Q} of Rational Numbers

The set \mathbb{Q} of Rational Numbers is the set of numbers that can be written as the ratio of two integers in \mathbb{Z} .

Definition 5.2 (*Algebraic number*)

A number x is called *algebraic* if it satisfies a polynomial equation:

$$c_n x^n + c_{n-1} x^{n-1} + \dots + c_1 x + c_0 = 0 \quad (5.2.1)$$

where $c_i \in \mathbb{Z}$ and $c_n \neq 0$.

Proposition 5.3 (*Rational \implies algebraic*)

All rational numbers are algebraic numbers.

Proof. Consider the rational number $x = \frac{m}{n} \in \mathbb{Q}$, where $m, n \in \mathbb{Z}$. Then, clearly it satisfies the equation:

$$nx - m = 0. \quad \blacksquare$$

Theorem 5.4 (*Rational Zeros Theorem*)

Assume $c_0 \dots c_n \in \mathbb{Z}$ and $x \in \mathbb{Q}$ satisfying the equation:

$$c_n x^n + c_{n-1} x^{n-1} + \dots + c_1 x + c_0 = 0 \quad (5.2.2)$$

where $c_n, c_0 \neq 0$. Let $x = \frac{c}{d}$ where $c, d \in \mathbb{Z}$ with no common factors and $d \neq 0$. Then c divides c_0 and d divides c_n .

Proof. We are given that:

$$\begin{aligned} c_n \left(\frac{c}{d}\right)^n + c_{n-1} \left(\frac{c}{d}\right)^{n-1} + \dots + c_1 \left(\frac{c}{d}\right) + c_0 &= 0 \\ c_n c^n + c_{n-1} c^{n-1} d + \dots + c_1 c d^{n-1} + c_0 d^n &= 0 \end{aligned}$$

Firstly, we solve for $c_0 d^n$ to obtain:

$$c_0 d^n = -c[c_n c^{n-1} + c_{n-1} c^{n-1} d + \dots + c_2 c d^{n-2} + c_1 d^{n-1}]$$

so c divides $c_0 d^n$. However, since c and d^n are coprime, we must have that c divides c_0 . Similarly, we solve for $c_n c^n$:

$$c_n c^n = -d[c_{n-1} c^{n-1} + c_{n-2} c^{n-2} d + \dots + c_1 c d^{n-2} + c_0 d^{n-1}]$$

so d divides $c_n c^n$. However, since c and d are coprime, we must have that d divides c_n as required. \blacksquare

Corollary Consider the equation:

$$x^n + c_{n-1} x^{n-1} + \dots + c_1 x + c_0 = 0. \quad (5.2.3)$$

By applying the Rational Zeros Theorem, all rational solutions must divide c_0 .

Example. Let us prove that $a = \sqrt{2 + \sqrt[3]{5}}$ is irrational.

Proof. We firstly note that a is algebraic, since:

$$\begin{aligned} a^2 &= 2 + \sqrt[3]{5} \\ (a^2 - 2)^3 &= 5 \\ a^6 - 6a^4 + 12a^2 - 13 &= 0 \end{aligned}$$

which gives the polynomial equation:

$$x^6 - 6x^4 + 12x^2 - 13 = 0 \quad (5.2.4)$$

By corollary 1.1.1 , the only possible rational solutions are $\pm 1, \pm 13$ which clearly don't satisfy (1.5). \blacksquare

5.3 The Set \mathbb{R} of Real Numbers

The following algebraic properties hold for a field:

A1. $a + (b + c) = (a + b) + c$ (addition associativity)

A2. $a + b = b + a$ (addition commutativity)

A3. $a + 0 = a$ (addition identity)

A4. $\forall a, \exists(-a)$ s.t. $a + (-a) = 0$ (addition inverse)

M1. $a(bc) = (ab)c$ (multiplication associativity)

M2. $ab = ba$ (multiplication commutativity)

M3. $a \cdot 1 = a$ (multiplication identity)

M4. $\forall a \neq 0, \exists a^{-1}$ s.t. $a \cdot a^{-1} = 1$ (multiplication inverse)

DL. $a(b + c) = ab + ac$ (distributivity)

The following ordering properties hold for an ordered field:

O1. $\forall a, b, a \leq b \vee b \leq a$ (multiplication inverse)

O2. $(a \leq b \wedge b \leq a) \implies a = b$

O3. $(a \leq b \wedge b \leq c) \implies a \leq c$

O4. $(a \leq b) \implies a + c \leq b + c$

O5. $(a \leq b \wedge 0 \leq c) \implies ac \leq bc$

Theorem 5.6 (Properties of fields)

The following are consequences of the field properties:

- (i) $a + c = b + c \implies a = b$
- (ii) $a \cdot 0 = 0$
- (iii) $(-a)b = -ab$
- (iv) $(-a)(-b) = ab$
- (v) $(ac = bc \wedge c \neq 0) \implies a = b$
- (vi) $ab = 0 \implies (a = 0 \vee b = 0)$

Proof. (i) $a + c = b + c \implies (a + c) + (-c) = (b + c) + (-c)$, using A1 we have that $a + [c + (-c)] = b + [c + (-c)] \implies a + 0 = b + 0$ by A4, so we finally have $a = b$ using A3.

- (ii) $a \cdot 0 = a \cdot (0 + 0) = a \cdot 0 + a \cdot 0$ where we used A3 and DL respectively. By (i) we conclude that $a \cdot 0 = 0$.
- (iii) $a + (-a) = 0 \implies ab + (-a)b = [a + (-a)] \cdot b = 0 \cdot b = 0 = ab + -(ab)$, so from (i) we have that $(-a)b = -(ab)$.

(iv) $(-a)(-b) + (-ab) = (-a)(-b) + (-a)b = (-a)[(-b) + b] = 0 = ab + (-ab)$, so by (i) we have $(-a)(-b) = ab$.

(v) Suppose $ac = bc \wedge c \neq 0$, then $a = a \cdot 1 = a(cc^{-1}) = (ac)c^{-1} = b(cc^{-1}) = b$

(vi) If $ab = 0$ and $b \neq 0$, then $0 = 0 \cdot b^{-1} = (ab)b^{-1} = a(bb^{-1}) = a \cdot 1 = a$

■

Theorem 5.7 (Properties of ordered fields)

The following are consequences of the properties of an ordered field:

- (i) $a \leq b \implies -b \leq -a$
- (ii) $(a \leq b \wedge c \leq 0) \implies bc \leq ac$
- (iii) $0 \leq a \wedge 0 \leq b \implies 0 \leq ab$
- (iv) $0 \leq a^2$
- (v) $0 < 1$
- (vi) $0 \leq a \implies 0 \leq a^{-1}$
- (vii) $0 < a < b \implies 0 < b^{-1} < a^{-1}$
- (viii) $0 \leq a, b \wedge p \in \mathbb{N} \implies (a < b \iff a^p < b^p)$

Proof. (i) Suppose $a \leq b$, then applying O4 with $c = (-a) + (-b)$, then $a + [(-a) + (-b)] \leq b + [(-a) + (-b)] \implies -b \leq -a$

(ii) if $a \leq b \wedge c \leq 0$, then $0 \leq -c$. So, applying O5 gives $a(-c) \leq -bc$, and using (i) we get $bc \leq ac$

(iii) This is a special case of O5 using $a = 0$.

(iv) For any a , $a \leq 0 \vee 0 \leq a$. In the first case, $a^2 \leq 0$ by (iii). In the latter case, $-a \leq 0 \implies (-a)(-a) = a^2 \leq 0$ using (i).

(v) Clearly, $0 \neq 1$. Indeed, consider $x \neq 0$, then $x \cdot 1 = x \cdot 0 = 0$ which is a contradiction. Applying (iii) with $a = 1$ gives the desired result.

(vi) Suppose $0 \leq a$ but $a^{-1} \leq 0 \implies -a^{-1} \geq 0$. Applying (iii) gives $0 \leq a(-a^{-1}) = -1 \implies 1 \leq 0$ which contradicts (v).

(vii) We multiply by $(a^{-1})(b^{-1}) > 0$ and find $0 < a < b \implies 0 < b^{-1} < a^{-1}$.

(viii) For positive integers p , we use the factor theorem:

$$b^p - a^p = (b - a) \underbrace{(b^{p-1} + b^{p-2}a + \dots + ba^{p-2} + a^{p-1})}_{>0} \quad (5.3.1)$$

but the term in brackets is positive definite, so it follows immediately that:

$$b - a > 0 \iff b^p - a^p > 0 \quad (5.3.2)$$



5.4 Absolute Value

Definition 5.8 (Absolute value and distance)

We define the *absolute value* of a as:

$$|a| = \begin{cases} a & \text{if } a \geq 0 \\ -a & \text{if } a \leq 0 \end{cases} \quad (5.4.1)$$

For two numbers a, b we define $\text{dist}(a, b) = |a - b|$ to be the *distance between a and b* .

We present some important properties of the absolute value:

Theorem 5.9 (Absolute value properties)

The following hold $\forall a, b \in \mathbb{R}$:

- (i) $|a| \geq 0$
- (ii) $|ab| = |a| \cdot |b|$
- (iii) $|a + b| \leq |a| + |b|$
- (iv) $|a - b| \geq ||a| - |b||$

Proof.

- (i) For $a \in \mathbb{R}$, $a \geq 0$ or $a \leq 0$, and since $|a| = \pm a$, it follows that $|a| \geq 0$.
- (ii) If $a \geq 0 \wedge b \geq 0$, then $ab \geq 0 \implies |a| \cdot |b| = ab = |ab|$. If $a \leq 0 \wedge b \leq 0$, then $ab \geq 0 \implies |a| \cdot |b| = (-a)(-b) = ab = |ab|$. If $a \geq 0 \wedge b \leq 0$, then $ab \leq 0 \implies |a| \cdot |b| = a(-b) = -ab = |ab|$. If $a \leq 0 \wedge b \geq 0$, then $ab \leq 0 \implies |a| \cdot |b| = (-a)b = -ab = |ab|$.
- (iii) By definition, we have $-|a| \leq a \leq |a|$ and $-|b| \leq b \leq |b|$, then O4 yields:

$$-|a| - |b| \leq a + b \leq |a| + |b|$$

so that

$$-(|a| + |b|) \leq a + b \leq |a| + |b|$$

which implies $a + b \leq |a| + |b|$ and $-(a + b) \leq |a| + |b|$. Since $|a + b| = \pm(a + b)$, it follows that $|a + b| \leq |a| + |b|$.

(iv) We proceed as follows:

$$|a - b| \geq ||a| - |b|| \iff (a - b)^2 \geq (|a| - |b|)^2 \quad (5.4.2)$$

$$\iff a^2 - 2ab + b^2 \geq a^2 - 2|a||b| + b^2 \quad (5.4.3)$$

$$\iff -2ab \geq -|2ab| \quad (5.4.4)$$

$$\iff ab \leq |ab| \quad (5.4.5)$$

which is true since $x \leq |x|$ for any real number x . ■

5.5 The Completeness Axiom

This axiom assures us that unlike \mathbb{Q} , \mathbb{R} has no gaps.

Definition 5.10 (Maximum and minimum)

Let S be a nonempty subset of \mathbb{R} .

- (i) If S contains a largest element s_0 s.t. $s_0 \in S \wedge s \leq s_0, \forall s \in S$, we write $s = \max S$.
- (ii) If S contains a smallest element we write it as $\min S$.

Example.

- (i) The set $\{r \in \mathbb{Q} | 0 \leq r \leq \sqrt{2}\}$ has a minimum, namely 0, but no maximum, since $\sqrt{2} \notin \mathbb{Q}$.
- (ii) Consider the set $\{n^{(-1)^n} | n \in \mathbb{N}\}$, which can be expanded into:

$$\{1, 2, \frac{1}{3}, 4, \frac{1}{5}, 6, \frac{1}{7}, \dots\}$$

which clearly has no maximum nor minimum. ◀

Definition 5.11 (Upper and lower bound)

Let S be a nonempty subset of \mathbb{R} :

- (i) If a real number M satisfies $s \leq M, \forall s \in S$, then M is called an upper bound of S .
- (ii) If a real number m satisfies $m \leq s, \forall s \in S$, then m is called a lower bound of S .
- (iii) A set S is bounded if it is bounded above and below i.e. $\exists m, M$ s.t. $S \subseteq [m, M]$.

Remark. Clearly, if a set S has a maximum, it is bounded above. Similarly, if it has a minimum, it is bounded below.

Definition 5.12 (Supremum and infimum)

Let S be a nonempty subset of \mathbb{R} .

- (i) If S is bounded above and has a least upper bound, then we call it the *supremum* of S , denoted by $\sup S$.
- (ii) If S is bounded above and has a least upper bound, then we call it the *infimum* of S , denoted by $\inf S$.

Remark. Observe that if S is bounded above, then $M = \sup S$ iff:

- (i) $s \leq M, \forall s \in S$
- (ii) $\forall M_1 < M, \exists s_1 \in S \text{ s.t. } s_1 > M_1$

Example.

- (a) If a set S has a maximum, then $\max S = \sup S$. Similarly, if a set S has a minimum, then $\min S = \inf S$.
- (b) We have $\inf\{n^{(-1)^n} : n \in \mathbb{N}\} = 0$.
- (c) The set $A = \{\frac{1}{n^2} : n \in \mathbb{N} \wedge n \geq 3\}$ is bounded. We have that $\sup A = \max A = \frac{1}{9}$ and the minimum does not exist, however $\inf A = 0$.
- (d) The set $B = \{r \in \mathbb{Q} : r^3 \leq 7\}$ is bounded above, but not below. It has no maximum, since $\sqrt[3]{7} \notin \mathbb{Q}$. However, $\sup B = \sqrt[3]{7}$ and $\inf B = -\infty$ since it has no minimum.
- (e) The set $C = \{m + n\sqrt{2} : m, n \in \mathbb{Z}\}$ is not bounded above or below, so it has no maximum or minimum. However, $\sup C = \infty$ and $\inf C = -\infty$.
- (f) The set $D = \{x \in \mathbb{R} : x^2 < 10\}$ is the open interval $(-\sqrt{10}, \sqrt{10})$. So, it is bounded above and below despite not having maximum and minimum. We have $\sup D = \sqrt{10}$ and $\inf D = -\sqrt{10}$.

**Theorem 5.13 (Completeness Axiom)**

Every nonempty subset S of \mathbb{R} that is bounded above has a least upper bound. In other words, $\sup S$ exists and is a real number.

Remark. Note that by this definition, the set of rationals \mathbb{Q} is incomplete, that is, it contains "gaps". Indeed, consider the set $A = \{r \in \mathbb{Q} : 0 \leq r \leq \sqrt{2}\}$, which is bounded above by $\frac{3}{2} \in \mathbb{Q}$ for example. If \mathbb{Q} were complete, then A would have a least upper bound that is rational, but such a number does not exist. **Corollary** Every nonempty subset S of \mathbb{R} that is bounded below has a greatest lower bound $\inf S$.

Proof. Let $-S$ be the set $\{-s : s \in S\}$ consisting of the negatives of S . Since S is bounded below, $\exists m \in \mathbb{R} \text{ s.t. } m \leq s \forall s \in S$. This implies that $-m \geq -s, \forall s \in S$, so since $-S$ is bounded above by $-m$, by the Completeness Axiom it must have a supremum. Let us now

prove that $\inf S = -\sup(-S)$. Let $s_0 = \sup(-S)$, then by definition $-s \leq s_0 \implies s \geq -s_0$, and since s_0 is the least upper bound of S , then if $t \geq -s$, $\forall s \in S$, then $t \geq s_0$. So, if $-t \leq s$, $\forall s \in S$, then $-t \leq -s_0$. These two conditions show that $-s_0$ is the greatest lower bound of S , so $\inf S = -\sup(-S)$, as required. \blacksquare

Theorem 5.14 (Archimedean Property)

If $a > 0$ and $b > 0$, then for some positive integer n , we have $na > b$.

Proof. Assume that Archimedean property fails, so there exists $a > 0$ and $b > 0$ such that $na \leq b$, $\forall n \in \mathbb{N} \implies b$ is an upper bound of $S = \{na : n \in \mathbb{N}\}$. Let $s_0 = \sup S$, and $a > 0 \implies s_0 - a < s_0$. Since s_0 is the supremum of S , $s_0 - a$ can't be an upper bound since it is smaller than s_0 , it follows that $\exists n_0 \in \mathbb{N}$ s.t. $s_0 - a < n_0 a \implies s_0 < (n_0 + 1)a \in S$ so s_0 is not an upper bound of S , which is a contradiction. \blacksquare

Theorem 5.15 (Dense ness of \mathbb{Q})

If $a, b \in \mathbb{R}$ and $a < b$, then $\exists r \in \mathbb{Q}$ s.t. $a < r < b$.

Proof. We wish to prove that:

$$a < r = \frac{m}{n} < b \implies an < m < bn$$

for some integers m, n . Since $b - a > 0$, by the Archimedean property, $\exists n \in \mathbb{N}$ s.t. $n(b - a) > 1$ \blacksquare

so it is evident that there is an integer m between an and bn since their difference is greater than 1.

Proposition 5.16 (Linearity of sup and inf)

For non empty bounded subsets A and B of \mathbb{R} , we have:

$$\sup(A + B) = \sup A + \sup B, \quad \inf(A + B) = \inf A + \inf B$$

Proof. Consider $x \in A + B \implies x = a + b$ for some $a \in A$, $b \in B$. It follows that $x \leq \sup A + \sup B \implies \sup(A + B) \leq \sup A + \sup B$. It remains to prove that $\sup(A + B) \geq \sup A + \sup B$. If one of the suprema is $+\infty$ (without loss of generality assume it is B), then taking some $a_0 \in A$, we have $\sup(A + B) \geq \sup(a_0 + B) = a_0 + \sup B = \infty = \sup A + \sup B$. If the sum of the suprema is finite, then we consider $\epsilon > 0$. Then $\exists a \in A$, $b \in B$, s.t. $a > \sup A - \frac{\epsilon}{2}$ and $b > \sup B - \frac{\epsilon}{2}$. It follows that $\sup(A + B) \geq a + b > \sup A + \sup B - \epsilon$ from which it follows that $\sup(A + B) \geq \sup A + \sup B$. \blacksquare

Unit D2: Sequences

6.1 Introduction to sequences

A sequence is a function whose domain a set of the form $\{n \in \mathbb{Z} : n \geq m\}$. The sequence is denoted by $(s)_{n=m}^{\infty}$ and the n th term in a sequence is denoted by s_n .

Example. Consider the sequence $(a_n)_{n=0}^{\infty}$ where $a_n = (-1)^n$, $n \geq 0$. We can then write the sequence as $(1, -1, 1, -1\dots)$, and has a set of values $\{-1, 1\}$. Note that the sequence contains infinite terms, but the set contains only two terms. ◀

Definition (*Monotonic sequence*)

A sequence (a_N) is said to be:

- (i) **constant** if $a_{n+1} = a_n$, for $n = 1, 2, 3\dots$
- (ii) **increasing (decreasing)** if $a_{n+1} \geq a_n$ ($a_{n+1} \leq a_n$) for $n = 1, 2, 3\dots$
- (iii) **strictly increasing** if $a_{n+1} > a_n$ ($a_{n+1} < a_n$) for $n = 1, 2, 3\dots$

If any of the above hold for (a_n) , then it is said to be **monotonic**.

Given a general sequence (a_n) , it is generally easier to show that:

- (i) $a_{n+1} - a_n \geq (\leq) 0 \implies (a_n)$ is increasing (decreasing)
- (ii) $a_{n+1} - a_n > (<) 0 \implies (a_n)$ is strictly increasing (strictly decreasing)
- (iii) $a_{n+1} - a_n = 0 \implies (a_n)$ is constant

Example. Consider the sequence $a_n = (n-1)(n-2)$, $n = 1, 2\dots$, which is monotonic increasing. Indeed, note that:

$$a_{n+1} - a_n = 2n - 2 \geq 0 \quad (6.1.1)$$

since $n \geq 1$. Note moreover that if $n \geq 2$, then (a_n) would be monotonic strictly increasing, since $2n - 2 > 0$.

Alternatively, it is also convenient to examine the quotient $\frac{a_{n+1}}{a_n}$.

Example. Consider the sequence $a_n = n + \frac{1}{n}$. Then:

$$\frac{a_{n+1}}{a_n} = \frac{n+1 + \frac{1}{n+1}}{n + \frac{1}{n}} = \frac{(n+1)^2}{n^2} \cdot \frac{n}{n+1} = \frac{n+1}{n} > 1 \quad (6.1.2)$$

since n is positive. It follows that a_n is monotonic strictly increasing. ◀

Definition (*Eventual properties*)

A sequence (a_n) eventually has a property if it satisfies the property for $n \geq n_0$ for some $n_0 \geq 1$.

Example. The sequence defined by $a_n = \frac{n^4}{4^n}$ is eventually decreasing. Indeed:

$$\frac{a_{n+1}}{a_n} = \frac{(n+1)^4}{n^4} \frac{4^n}{4^{n+1}} = \frac{1}{4} \left(\frac{n+1}{n} \right)^4 < 1 \implies 1 + \frac{1}{n} < \sqrt{2} \implies \frac{1}{\sqrt{2}-1} < n \quad (6.1.3)$$

so the sequence eventually decreases, more specifically for $n \geq 3$. ◀

6.2 Convergence of sequences

Definition 6.1 (*Sequence convergence*)

A sequence (s_n) of real numbers converges to s (i.e. $\lim_{n \rightarrow \infty} s_n = s$) provided that:

$$\forall \epsilon > 0, \exists N \in \mathbb{R} \text{ s.t. } n > N \implies |s_n - s| < \epsilon$$

A sequence that does not converge to a real number is divergent.

Example. Consider the sequence $s_n = \frac{3n+1}{7n-4} = \frac{3+\frac{1}{n}}{7-\frac{4}{n}}$ then clearly for large values of n , the series should converge to $\frac{3}{7}$. Indeed, by definition 7.1, $\lim_{n \rightarrow \infty} s_n = \frac{3}{7}$ means that:

$$\forall \epsilon > 0, \exists N \text{ s.t. } n > N \implies \left| \frac{3n+1}{7n-4} - \frac{3}{7} \right| < \epsilon$$

As ϵ varies, getting smaller and smaller, N gets bigger and bigger, so in the end for $n > N$, so a very large value of n , the difference between s_n and s becomes very very small, which intuitively makes sense. ◀

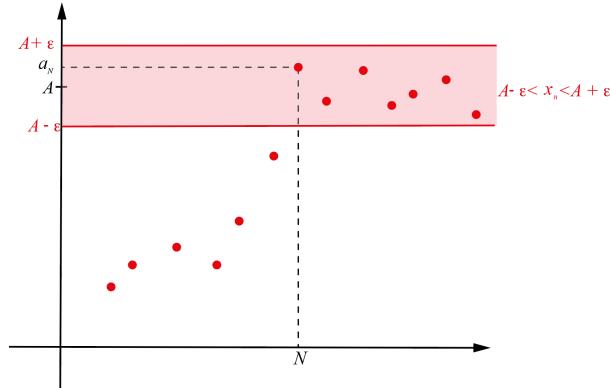


Figure 6.1. Geometrical interpretation of the epsilon-delta definition

Remark. Finally, it must be noted that limits are unique, so:

$$\lim_{n \rightarrow \infty} s_n = s \wedge \lim_{n \rightarrow \infty} s_n = t \implies s = t$$

Indeed, the first implies that:

$$\exists N_1 \text{ s.t. } n > N_1 \implies |s_n - s| < \frac{\epsilon}{2}$$

and the second implies that:

$$\exists N_2 \text{ s.t. } n > N_2 \implies |s_n - t| < \frac{\epsilon}{2}$$

for some $\epsilon > 0$. For $n > \max\{N_1, N_2\}$ (this allows us to use both conditions of convergence) the triangle inequality shows:

$$|s - t| = |(s - s_n) + (s_n - t)| \leq |s - s_n| + |s_n - t| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

for all $\epsilon > 0$ and thus $|s - t| = 0 \implies s = t$ as required. Geometrically, this argument corresponds to showing that for $n > \max\{N_1, N_2\}$, the terms of the sequence can't belong to both $\{y_1 : |y_1 - s| < \frac{\epsilon}{2}\}$ and $\{y_2 : |y_2 - t| < \frac{\epsilon}{2}\}$ since the two don't intersect for sufficiently small ϵ . ■

n	$s_n = \frac{3n+1}{7n-4}$	$ s_n - \frac{3}{7} $
2	0.7000	0.2714
3	0.5882	0.1597
5	0.5161	0.0876
40	0.4384	0.0098
400	0.4295	0.0010

6.3 Formal Proofs of Limit Theorems

Example. Let us prove that $\lim_{n \rightarrow \infty} \frac{3n+1}{7n-4} = \frac{3}{7}$.

Discussion: we consider an arbitrary $\epsilon > 0$ and show that $\exists N$ such that $n > N \implies |\frac{3n+1}{7n-4} - \frac{3}{7}| < \epsilon$. Thus, we want:

$$|\frac{19}{7(7n-4)}| < \epsilon$$

and since $7n - 4 > 0$ since n is positive, we drop the absolute value and write:

$$\frac{19}{7(7n-4)} < \epsilon \implies \frac{19}{7\epsilon} < 7n-4 \implies \frac{19}{49\epsilon} + \frac{4}{7} < n$$

so we put $N = \frac{19}{49\epsilon} + \frac{4}{7}$.

Proof: let $\epsilon > 0$ and $N = \frac{19}{49\epsilon} + \frac{4}{7}$. Then

$$n > N \implies n > \frac{19}{49\epsilon} + \frac{4}{7} \implies 7n - 4 > \frac{19}{7\epsilon} \implies \epsilon > \frac{19}{7(7n-4)}$$

Thus $n > N \implies |\frac{3n+1}{7n-4} - \frac{3}{7}| < \epsilon$ which proves that $\lim_{n \rightarrow \infty} \frac{3n+1}{7n-4} = \frac{3}{7}$ as required.

◀

Example. Let us prove that $\lim_{n \rightarrow \infty} \frac{4n^3+3n}{n^3-6} = 4$.

Discussion: For each $\epsilon > 0$, we need the following inequality to hold:

$$|\frac{4n^3+3n}{n^3-6} - 4| < \epsilon \implies |\frac{3n+24}{n^3-6}| < \epsilon$$

Note that it is very hard in this case to solve for n , so instead of finding the smallest N such that $n > N$ implies that $|\frac{3n+24}{n^3-6}| < \epsilon$, we will use estimates. Note that $3n+24 \leq 27n$ for $n > 1$ and $n^3 - 6 \geq \frac{n^3}{2}$ for $n > 2$. So:

$$|\frac{3n+24}{n^3-6}| \leq \frac{27n}{\frac{1}{2}n^3} < \epsilon \implies n > \sqrt{\frac{54}{\epsilon}}$$

if $n > 2$.

Proof: let $\epsilon > 0$ and $N = \max\{2, \sqrt{\frac{54}{\epsilon}}\}$. Then:

$$n > N \implies n > \sqrt{\frac{54}{\epsilon}} \implies \frac{27n}{n^3/2} < \epsilon$$

Since for $n > 2$, $3n + 24 \leq 27n$ and $n^3 - 6 \geq \frac{n^3}{2}$, we find that:

$$\left| \frac{3n + 24}{n^3 - 6} \right| \leq \frac{27n}{n^3/2} < \epsilon$$

and hence

$$\left| \frac{4n^3 + 3n}{n^3 - 6} - 4 \right| < \epsilon$$

as required. \blacktriangleleft

Example. Let (s_n) be a sequence of non-negative real numbers such that $\lim_{n \rightarrow \infty} s_n = s$. Then, $\lim_{n \rightarrow \infty} \sqrt{s_n} = \sqrt{s}$.

Proof.

Case I: let $\epsilon > 0$ and $s > 0$, since $\lim_{n \rightarrow \infty} s_n = s$,

$$\exists N \text{ s.t. } n > N \implies |s_n - s| < \sqrt{s}\epsilon$$

so:

$$\exists N \text{ s.t. } n > N \implies |\sqrt{s_n} - \sqrt{s}| = \frac{|s_n - s|}{\sqrt{s_n} + \sqrt{s}} \leq \frac{|s_n - s|}{\sqrt{s}} < \frac{\sqrt{s}}{\sqrt{s}}\epsilon = \epsilon$$

as desired. \blacktriangleleft

Case II: if $s = 0$, let $\epsilon > 0$ so that:

$$\exists N \text{ s.t. } n > N \implies |s_n| < \epsilon^2$$

Hence, $\sqrt{s_n} < \epsilon$ for $n > N$ and thus:

$$|\sqrt{s_n} - 0| < \epsilon \implies \lim_{n \rightarrow \infty} \sqrt{s_n} = s = 0$$

as desired. \blacktriangleleft

Example. Let (s_n) be a convergent sequence of non-zero real numbers such that $\lim_{n \rightarrow \infty} s_n = s \neq 0$. Then $\inf\{|s_n| : n \in \mathbb{N}\} > 0$.

Discussion: The result has the geometric interpretation that all terms of the sequence are "close" to s and therefore not "close" to 0. The proof will involve three steps. First, we show that there exists N such that all terms of the sequence after s_N are all greater than $\frac{|s|}{2}$ by using the triangle inequality. This result shows that the terms of the sequence after s_N are all "close" to s , with a maximum distance of $\frac{|s|}{2}$. We then take the minimum of $\frac{|s|}{2}$, and $|s|_N$, and show that it is positive so that $\inf\{|s_n| : n \in \mathbb{N}\} > 0$ follows directly.

Proof.

Let $\epsilon = \frac{|s|}{2} > 0$, since $\lim_{n \rightarrow \infty} s_n = s$:

$$\exists N \text{ s.t. } n > N \implies |s_n - s| \leq \frac{|s|}{2}$$

We must have that:

$$\exists N \text{ s.t. } n > N \implies |s_n| \geq \frac{|s|}{2}$$

since:

$$|s| = |s - s_n + s_n| \leq |s - s_n| + |s_n| \leq \frac{|s|}{2} + |s_n| \implies |s_n| \geq \frac{|s|}{2}.$$

Setting:

$$m = \min\left\{\frac{|s|}{2}, |s_1|, \dots, |s_n|\right\}$$

In view of this result, $|s_n| \geq m$, so $\inf\{|s_n| : n \in \mathbb{N}\} > 0$ as required.



6.4 Null sequences

Definition (Null sequence)

A sequence (a_n) is said to be null if it converges to zero, that is, if:

$$\forall \epsilon > 0, \exists N \in \mathbb{R} \text{ s.t. } n > N \implies |a_n| < \epsilon \quad (6.4.1)$$

Example. Consider the sequence

$$a_n = \frac{(-1)^n}{n^4 + 1}, \quad n = 1, 2, \dots \quad (6.4.2)$$

Suppose $0 < \epsilon < 1$, then:

$$|a_n| = \frac{1}{n^4 + 1} < \epsilon \iff n^4 > \frac{1}{\epsilon} - 1 \iff n > \left(\frac{1}{\epsilon} - 1\right)^{\frac{1}{4}} = N \quad (6.4.3)$$

so for a given $0 < \epsilon < 1$, $|a_n| < \epsilon$ provided that $n > N$ where $N = \left(\frac{1}{\epsilon} - 1\right)^{\frac{1}{4}}$.

If instead $\epsilon \geq 1$, then:

$$|a_n| = \frac{1}{n^4 + 1} < \epsilon \iff n^4 > \frac{1}{\epsilon} - 1 \quad (6.4.4)$$

but since $\epsilon \geq 1 \implies \frac{1}{\epsilon} \leq 1$, the LHS is non-positive, so the above inequality is true for all $n > N$ where $N = 1$.



Theorem (Power rule of null sequences)

If (a_n) is a null, non-negative sequence for $n = 1, 2, \dots$, then the sequence (a_n^p) given by a_n^p , with $p \in \mathbb{R}$ for all $n = 1, 2, \dots$ is also null.

Proof. The sequence (a_n) is null, therefore for each positive $\epsilon^{1/p}$ there exists N such that:

$$a_n < \epsilon^{1/p}, \quad \forall n > N \quad (6.4.5)$$

so that for all positive ϵ :

$$a_n^p < \epsilon, \quad \forall n > N \quad (6.4.6)$$

so $\lim_{n \rightarrow \infty} a_n^p = 0$ as desired. ■

Example. Consider the sequence $a_n = \frac{1}{n}$ for $n = 1, 2, \dots$. Then a_n is null, since for all $\epsilon > 0$:

$$|a_n| = \frac{1}{n} < \epsilon \iff \epsilon < n, \quad (6.4.7)$$

so that $|a_n| \leq \epsilon$ for all $n > N = \epsilon$. Hence $\lim_{n \rightarrow \infty} a_n = 0$. Applying the power rule to a_n we then find that $\frac{1}{n^p}$ is also null. ◀

Proposition (Limit theorems for null sequences)

Let (a_n) , (b_n) and (c_n) be null sequences, and let $\alpha \in \mathbb{R}$. Then:

- (i) (αa_n) is null
- (ii) $(a_n + b_n)$ is null
- (iii) $(a_n b_n)$ is null
- (iv) $\left(\frac{a_n}{b_n}\right)$ is null

Proof. Immediate application of the Limit theorems (which will be shown in the next section) will give the desired results. ■

Proposition (Standard null sequences)

The following sequences are all null:

- (i) $\left(\frac{1}{n^p}\right)$ for $p > 0$
- (ii) (c^n) for $|c| < 1$
- (iii) $(n^p c^n)$ for $p > 0, |c| < 1$
- (iv) $\left(\frac{c^n}{n!}\right)$ for $c \in \mathbb{R}$
- (v) $\left(\frac{n^p}{n!}\right)$ for $p > 0$.

Proof. (i) It has been proven in the example preceding the limit theorems.

- (ii) Note that a_n is null $\iff |a_n|$ is also null, so it suffices to prove that $|c^n| = |c|^n$ is null (see the Theorem on the convergence of absolute values). In other words, we can limit ourselves to c non-negative.

Suppose that $c = 0$, then the nullity is trivial. Suppose that $0 < c < 1$, then:

$$c = \frac{1}{a+1}, \quad a > 0 \quad (6.4.8)$$

We may apply bernoulli's inequality:

$$c^n = \frac{1}{(a+1)^n} \leq \frac{1}{na+1} \leq \frac{1}{na} \quad (6.4.9)$$

Since $\frac{1}{n}$ is null, it follows from the limit theorem (ii) that $\frac{1}{na}$ is also null. Applying the squeeze theorem then, we finally find that c^n is also null.

- (iii) Once again assume that $0 < c < 1$, then for some $a > 0$:

$$c = \frac{1}{a+1} \quad (6.4.10)$$

We begin by proving that (nc^n) is null. By the binomial theorem:

$$(1+a)^n \geq 1 + na + \frac{1}{2}n(n-1)a^2 \geq \frac{1}{2}n(n-1)a^2, \quad n = 2, 3, \dots \quad (6.4.11)$$

it follows that

$$nc^n = \frac{n}{(a+1)^n} \leq \frac{n}{\frac{1}{2}n(n-1)a^2} = \frac{2}{a^2(n-1)}, \quad n = 2, 3, \dots \quad (6.4.12)$$

The sequence defined by:

$$b_n = \frac{2}{a^2(n-1)}, \quad n = 2, 3, \dots \quad (6.4.13)$$

is equivalent to:

$$b_n = \frac{2}{a^2n}, \quad n = 1, 2, \dots \quad (6.4.14)$$

which is null by the limit theorems. By the squeeze rule, it follows that (nc^n) is also null.

For $(n^p c^n)$, where $p > 0$ and $0 < c < 1$ we write:

$$n^p c^n = (nd^n)^p, \quad n = 1, 2, \dots \quad (6.4.15)$$

where $0 < d = c^{1/p} < 1$. We have shown that (nd^n) is null, so that by the power rule $(n^p c^n)$ is also null.

- (iv) Again, as in the case of (ii) we may consider only positive values of c . Let us choose

an integer m such that $m + 1 > c$, so that for $n > m + 1$:

$$\frac{c^n}{n!} = \prod_{k=1}^n \frac{c}{k} \leq \frac{c}{n} \prod_{k=1}^m \frac{c}{k} = \frac{c^m}{m!} \cdot \frac{c}{n} \quad (6.4.16)$$

Since $\frac{1}{n}$ is null, we have that $\frac{c^m}{m!} \cdot \frac{c}{n}$ too is null, since $\frac{c^{m+1}}{m!}$ is just a constant. By the squeeze rule, it follows that $\frac{c^n}{n!}$ is null, as desired.

(v) We can write:

$$\frac{n^p}{n!} = \frac{n^p}{2^n} \cdot \frac{2^n}{n!} \quad (6.4.17)$$

so that $\left(\frac{n^p}{n!}\right)$ is null by the limit theorem (iii). ■

Example. Consider the sequence described by

$$a_n = \frac{1}{3n^4(2n-1)^{1/3}}, \quad n = 1, 2, \dots \quad (6.4.18)$$

We can rewrite the terms of this sequence as:

$$a_n = \frac{1}{3} \cdot \frac{1}{n^4} \cdot \left(\frac{(-1)^n}{n^4+1}\right)^{1/3} \quad (6.4.19)$$

We know that $\frac{1}{n^4}$ is null, since it is the first standard null sequence with $p = 1$. Moreover, we have determined previously that $\frac{(-1)^n}{n^4+1}$ is null, so that by the power rule, $\left(\frac{(-1)^n}{n^4+1}\right)^{1/3}$ is also null. Finally, we exploit the product rule to conclude that a_n is indeed null. ◀

Theorem (Squeeze rule for null sequences)

If (b_n) is a null sequence of non-negative terms and:

$$|a_n| \leq b_n, \quad \forall n = 1, 2, \dots \quad (6.4.20)$$

then (a_n) is null.

Proof. We apply the squeeze rule (which will be proven in the next section) with:

$$-b_n \leq a_n \leq b_n, \quad \forall n = 1, 2, \dots \quad (6.4.21)$$

then since $(-b_n)$ and (b_n) both converge to 0, then $\lim_{n \rightarrow \infty} a_n = 0$ as well. ■

Example. Consider the sequence:

$$a_n = \frac{\sin(n^2)}{n^2 + 2^n}, \quad n = 1, 2, \dots \quad (6.4.22)$$

Then:

$$\left| \frac{\sin(n^2)}{n^2 + 2^n} \right| \leq \left| \frac{1}{n^2 + 2^n} \right| \quad (6.4.23)$$

Since both n^2 and 2^n are positive:

$$\frac{1}{n^2 + 2^n} < \frac{1}{n^2} \quad (6.4.24)$$

so:

$$\left| \frac{\sin(n^2)}{n^2 + 2^n} \right| \leq \frac{1}{n^2} \quad (6.4.25)$$

However, $\frac{1}{n^2}$ is one of the standard null sequences, so we find by the Squeeze rule that a_n is also null.



6.5 Limit theorems for Convergent sequences

Definition (Bounded and unbounded sequence)

A sequence (a_n) is **bounded** if $\exists M$ such that:

$$|a_n| \leq M, \quad n = 1, 2, \dots \quad (6.5.1)$$

and is **unbounded** otherwise

Example. The sequence $a_n = \frac{2n+1}{n}$ for $n = 1, 2, \dots$ is bounded. Indeed, for all natural numbers n we find that:

$$a_n = \frac{2n+1}{n} = 2 + \frac{1}{n} \leq 2 + 1 = 3 \quad (6.5.2)$$

since $\frac{1}{n} \leq 1$.



Example. Consider the sequence $b_n = (-1)^n n$, for $n = 1, 2, \dots$, then suppose there exists some M such that:

$$|b_n| = |n| = n < M \quad (6.5.3)$$

for all $n = 1, 2, \dots$, which is clearly a contradiction since \mathbb{N} has no maximum. Consequently, b_n is unbounded.

◀

Proposition (Boundedness of convergent sequences)

Convergent sequences are bounded, and unbounded sequences are divergent.

Proof. Suppose (s_n) is a convergent sequence with $\lim_{n \rightarrow \infty} s_n = s$. Then we can apply the Definition 6.1 with $\epsilon = 1$ to find:

$$n > N \implies |s_n - s| < 1 \implies |s_n| < 1 + |s| \quad (6.5.4)$$

where we use the triangle inequality. Let us now define $M = \max\{|s| + 1, |s_1|, \dots, |s_N|\}$ so that $|s_n| \leq M$ for all n , proving the boundedness of convergent sequences.

The converse follows immediately. ■

Remark. Note that it does not suffice to show that $|s_n| < 1 + |s|$, because this only proves that the terms of the sequence after s_N are bounded. We must introduce M in order to prove boundedness for the initial N terms.

Also note that the choice $\epsilon = 1$ was completely arbitrary.

Example. Consider the sequence $a_n = n^{(-1)^n}$ for $n = 1, 2, \dots$. Then, for n even:

$$|a_n| = n^{(-1)^n} = n \quad (6.5.5)$$

which is unbounded, since \mathbb{N} is also unbounded. Instead for n odd:

$$|a_n| = \frac{1}{n} < 1 \quad (6.5.6)$$

which is bounded. Consequently, a_n overall is unbounded, and thus divergent.

◀

Example. Consider the sequence $a_n = \frac{n^2+n}{n^2+1}$ for $n = 1, 2, \dots$. Then

$$\frac{n^2+n}{n^2+1} \leq \frac{n^2+n^2}{n^2} = 2 \quad (6.5.7)$$

so the sequence is bounded. Moreover,

$$\lim_{n \rightarrow \infty} \frac{n^2 + n}{n^2 + 1} = \lim_{n \rightarrow \infty} \frac{1 + \frac{1}{n}}{1 + \frac{1}{n^2}} = 1 \quad (6.5.8)$$

so it also converges.



Theorem 6.3 (Limit theorems)

Let $(s_n), (t_n)$ sequences converging to s, t respectively, and let $\alpha \in \mathbb{R}$. Then:

- (i) $\lim_{n \rightarrow \infty} (ks_n) = k \lim_{n \rightarrow \infty} s_n$
- (ii) $\lim_{n \rightarrow \infty} (s_n + t_n) = \lim_{n \rightarrow \infty} s_n + \lim_{n \rightarrow \infty} t_n$
- (iii) $\lim_{n \rightarrow \infty} (s_n t_n) = (\lim_{n \rightarrow \infty} s_n)(\lim_{n \rightarrow \infty} t_n)$
- (iv) $\lim_{n \rightarrow \infty} \left(\frac{t_n}{s_n} \right) = \frac{\lim_{n \rightarrow \infty} t_n}{\lim_{n \rightarrow \infty} s_n}$ if $\lim_{n \rightarrow \infty} s_n \neq 0$ and $s_n \neq 0$ for all n

Proof.

- (i) If $k = 0$ then the result is trivial. If we assume that $k \neq 0$, and we let $\epsilon > 0$, then since $\lim_{n \rightarrow \infty} s_n = s$ for some $s \in \mathbb{R}$ there exists N such that:

$$n > N \implies |s_n - s| < \frac{\epsilon}{|k|} \quad (6.5.9)$$

Then:

$$n > N \implies |ks_n - ks| < \epsilon \quad (6.5.10)$$

showing that $\lim_{n \rightarrow \infty} (ks_n) = k \lim_{n \rightarrow \infty} s_n$ as desired.

- (ii) Let $\epsilon > 0$, we need to show that for appropriately large n :

$$|s_n + t_n - (s + t)| < \epsilon \quad (6.5.11)$$

Note however that since $\lim_{n \rightarrow \infty} s_n = s$, there exists N_1 such that:

$$n > N_1 \implies |s_n - s| < \frac{\epsilon}{2} \quad (6.5.12)$$

and similarly for t_n :

$$n > N_2 \implies |t_n - t| < \frac{\epsilon}{2} \quad (6.5.13)$$

so that, letting $N = \max\{N_1, N_2\}$

$$n > N \implies |s_n - s| + |t_n - t| < \epsilon \quad (6.5.14)$$

Finally, we make use of the triangle inequality:

$$n > N \implies |s_n + t_n - (s + t)| \leq |s_n - s| + |t_n - t| < \epsilon \quad (6.5.15)$$

as desired.

(iii) Firstly note that:

$$|s_n t_n - st| = |s_n t_n - s_n t + s_n t - st| \leq |s_n| \cdot |t_n - t| + |t| \cdot |s_n - s| \quad (6.5.16)$$

Let $\epsilon > 0$, by theorem 6.2 s_n must be bounded, so we can find $M > 0$ such that $|s_n| \leq M$ for all n . Since t_n converges:

$$n > N_1 \implies |t_n - t| < \frac{\epsilon}{2M} \quad (6.5.17)$$

and similarly for s_n :

$$n > N_2 \implies |s_n - s| < \frac{\epsilon}{2(|t| + 1)} \quad (6.5.18)$$

where we used $|t|+1$ since it could be the case that $|t| = 0$. Suppose $N = \max\{N_1, N_2\}$ then:

$$n > N \implies |s_n t_n - st| \leq M \frac{\epsilon}{2M} + |t| \frac{\epsilon}{2(|t| + 1)} < \epsilon \quad (6.5.19)$$

as desired, $\lim_{n \rightarrow \infty} (s_n t_n) = (\lim_{n \rightarrow \infty} s_n)(\lim_{n \rightarrow \infty} t_n)$.

(iv) We begin by proving the following lemma:

Lemma. If s_n converges to $s \neq 0$ and $s_n \neq 0$ for all n , then $\frac{1}{s_n}$ converges to $\frac{1}{s}$.

To prove this lemma, let $\epsilon > 0$. In the last example of the previous section, we showed that there exists $m > 0$ such that $|s_n| \geq m$ for all n . The convergence of s_n means that there exists N so that:

$$n > N \implies |s - s_n| < \epsilon m |s| \quad (6.5.20)$$

Then $n > N$ implies:

$$\left| \frac{1}{s_n} - \frac{1}{s} \right| = \frac{|s - s_n|}{|s_n s|} \leq \frac{|s - s_n|}{m |s|} \leq \epsilon \quad (6.5.21)$$

as desired.

We may finally use this lemma to prove the main result:

$$\lim_{n \rightarrow \infty} \frac{t_n}{s_n} = \lim_{n \rightarrow \infty} \frac{1}{s_n} t_n = \frac{t}{s} \quad (6.5.22)$$

as desired. ■

Remark. Note that we also impose the condition $s_n \neq 0$ in order for the reciprocal sequence $\frac{1}{s_n}$ to be well-defined.

Example. Consider the sequence:

$$a_n = \frac{n^2 + 2^n}{3^n + n^3} \quad (6.5.23)$$

We can factorize out 2^n on the numerator and 3^n on the denominator to find that:

$$a_n = \frac{2^n}{3^n} \cdot \frac{\frac{n^2}{2^n} + 1}{\frac{n^3}{3^n} + 1} = \left(\frac{2}{3}\right)^n \cdot \frac{\frac{n^2}{2^n} + 1}{\frac{n^3}{3^n} + 1} \quad (6.5.24)$$

Hence:

$$\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} \left[\left(\frac{2}{3}\right)^n \cdot \left(\frac{\frac{n^2}{2^n} + 1}{\frac{n^3}{3^n} + 1} \right) \right] \quad (6.5.25)$$

$$= \lim_{n \rightarrow \infty} \left(\frac{2}{3}\right)^n \cdot \frac{\lim_{n \rightarrow \infty} \frac{n^2}{2^n} + 1}{\lim_{n \rightarrow \infty} \frac{n^3}{3^n} + 1} \quad (6.5.26)$$

$$= \lim_{n \rightarrow \infty} \left(\frac{2}{3}\right)^n \quad (6.5.27)$$

$$= 0 \quad (6.5.28)$$

so a_n is a null sequence.



Theorem (Squeeze rule)

Let (a_n) , (b_n) and (c_n) be convergent sequences such that:

- (i) $b_n \leq a_n \leq c_n$ for $n = 1, 2, \dots$
- (ii) $\lim_{n \rightarrow \infty} b_n = \lim_{n \rightarrow \infty} c_n = l$

Then $\lim_{n \rightarrow \infty} a_n = l$.

Proof. From condition (ii) we have that for all $\epsilon > 0$ then there exists N_1 and N_2 such that

$$l - \epsilon < b_n < l + \epsilon, \quad \forall n > N_1 \quad (6.5.29)$$

$$l - \epsilon < c_n < l + \epsilon, \quad \forall n > N_2 \quad (6.5.30)$$

Then, it follows from condition (i) that:

$$l - \epsilon < b_n \leq a_n \leq c_n < l + \epsilon, \quad \forall n > N \quad (6.5.31)$$

where $N = \max\{N_1, N_2\}$, so that $\lim_{n \rightarrow \infty} a_n = l$ as desired. ■

Example. Consider the sequence $a_n = n^{1/n}$. Using the binomial theorem for $n \geq 2$ and $x \geq 0$:

$$(1+x)^n \geq \frac{n(n-1)}{2}x^2 \quad (6.5.32)$$

since all the other terms in the binomial expansion are positive. Substituting $x = \sqrt{\frac{2}{n-1}}$ we find that:

$$\left(1 + \sqrt{\frac{2}{n-1}}\right)^n \geq \frac{n(n+1)}{2} \frac{2}{n-1} = n \quad (6.5.33)$$

implying that

$$n^{1/n} \leq 1 + \sqrt{\frac{2}{n-1}}, \quad n = 2, 3, \dots \quad (6.5.34)$$

If we define b_n to be:

$$b_1 = 1 \quad (6.5.35)$$

$$b_n = 1 + \sqrt{\frac{2}{n-1}}, \quad n = 2, 3, \dots \quad (6.5.36)$$

then we see that $a_n \leq b_n$ for $n = 1, 2, 3, \dots$. Moreover, $\lim_{n \rightarrow \infty} b_n = 1$. Indeed for all $\epsilon > 0$:

$$|b_n - 1| < \epsilon \iff n > \frac{2}{\epsilon^2} + 1 = N > 1 \quad (6.5.37)$$

so there exists some N such that $|b_n - 1| < \epsilon$. It is important that $N > 1$, since otherwise it can be the case for some $n > N$ that $b_n - 1 = 0$.

It follows from the squeeze rule that:

$$\lim_{n \rightarrow \infty} n^{1/n} = 1 \quad (6.5.38)$$



Proposition (Limit inequality)

If $\lim_{n \rightarrow \infty} a_n = l$, $\lim_{n \rightarrow \infty} b_n = m$ and:

$$a_n \leq b_n, \quad \forall n = 1, 2, \dots \quad (6.5.39)$$

then $l \leq m$.

Proof. Suppose that $\lim_{n \rightarrow \infty} a_n = l$, $\lim_{n \rightarrow \infty} b_n = m$ and $a_n \leq b_n$ for $n = 1, 2, \dots$. If $l > m$ then:

$$\lim_{n \rightarrow \infty} (a_n - b_n) = l - m > 0 \quad (6.5.40)$$

so that:

$$a_n - b_n > \frac{1}{2}(l - m) \quad (6.5.41)$$

Indeed, if we take $\epsilon = \frac{l}{2}$ then it follows that:

$$|(a_n - b_n) - (l - m)| < \frac{l}{2} \implies -\frac{l-m}{2} < (a_n - b_n) - (l - m) < \frac{l-m}{2} \quad (6.5.42)$$

$$\implies \frac{l-m}{2} < a_n, \forall n > N \quad (6.5.43)$$

for some N . However, we also have that $a_n - b_n \leq 0$ so that:

$$l - m \leq 0 \implies l \leq m \quad (6.5.44)$$

a contradiction.

Proposition (Limit uniqueness)

The limit of a sequence is unique.

Suppose $\lim_{n \rightarrow \infty} a_n = l$ and $\lim_{n \rightarrow \infty} a_n = m$, then using the Limit inequality rule we get that $l \leq m$ and $m \leq l$, giving $l = m$ as desired. ■

Theorem (Convergence of absolute value)

If $\lim_{n \rightarrow \infty} a_n = l$ then $\lim_{n \rightarrow \infty} |a_n| = |l|$.

Proof. Using the triangle inequality:

$$||a_n| - |l|| \leq |a_n - l| \quad (6.5.45)$$

Therefore, for all $\epsilon > 0$, there exists N such that:

$$|a_n - l| \leq \epsilon, \forall n > N \quad (6.5.46)$$

implying that:

$$||a_n| - |l|| \leq \epsilon, \forall n > N \quad (6.5.47)$$

It follows that $\lim_{n \rightarrow \infty} |a_n| = |l|$ as expected. ■

6.6 Divergent sequences

Definition (Infinite limit)

A sequence (a_n) tends to infinity if:

$$\forall M > 0, \exists N \text{ s.t. } a_n > M, \forall n > N \quad (6.6.1)$$

We write that $a_n \rightarrow \infty$.

The sequence (a_n) tends to minus infinity if:

$$-a_n \rightarrow \infty \quad (6.6.2)$$

Theorem (Reciprocal rule)

If (a_n) is eventually positive, and $\left(\frac{1}{a_n}\right)$ is a null sequence then $a_n \rightarrow \infty$.

Proof. Let $M > 0$, then since a_n is eventually positive:

$$a_n > 0, \forall n > N_1 \quad (6.6.3)$$

Moreover, $\frac{1}{a_n}$ is null so that taking $\epsilon = \frac{1}{M}$:

$$\left| \frac{1}{a_n} \right| < \frac{1}{M_*}, \forall n > N_2 \quad (6.6.4)$$

Taking $N = \max\{N_1, N_2\}$ then:

$$0 < \frac{1}{a_n} < \frac{1}{M}, \forall n > N \quad (6.6.5)$$

implying:

$$a_n > M, \forall n > N \quad (6.6.6)$$

as desired. ■

Example. Consider the sequence $a_n = n! - 10^n$ for $n = 1, 2, \dots$. The dominant term is $n!$, so let us write:

$$a_n = n! \left(1 - \frac{10^n}{n!}\right), n = 1, 2, \dots \quad (6.6.7)$$

Therefore, a_n is indeed eventually positive, since $10^n < n!$ is true for $n \geq 25$. Then:

$$\lim_{n \rightarrow \infty} \frac{1}{a_n} = \lim_{n \rightarrow \infty} \frac{1}{n!} \frac{1}{1 - \frac{10^n}{n!}} = 0 \cdot \frac{0}{1 - 0} = 0 \quad (6.6.8)$$

By the reciprocal rule we see that $a_n \rightarrow \infty$. ◀

Proposition (Properties of diverging sequences)

If (a_n) and (b_n) both tend to infinity, then:

- (i) $(a_n + b_n)$ tends to infinity

- (ii) (αa_n) tends to infinity
- (iii) $(a_n b_n)$ tends to infinity

Theorem (Squeeze theorem for sequences tending to ∞)

If (b_n) tends to infinity and $a_n \geq b_n$ for $n = 1, 2, \dots$ then (a_n) tends to infinity.

Example. Consider the sequence $a_n = \frac{2^n}{n} + 5n^9$, $n = 1, 2, \dots$. Then let us define $b_n = \frac{2^n}{n}$ and $c_n = n^9$ so that:

$$a_n = b_n + 5c_n \quad (6.6.9)$$

Now $b_n \rightarrow \infty$, since it is eventually positive for $n \geq 1$ and:

$$\lim_{n \rightarrow \infty} \frac{1}{b_n} = \lim_{n \rightarrow \infty} n \left(\frac{1}{2} \right)^n = 0 \quad (6.6.10)$$

where we used the standard null sequence (iii) with $p = 1$.

Similarly, $c_n \rightarrow \infty$ since:

$$\lim_{n \rightarrow \infty} \frac{1}{c_n} = \lim_{n \rightarrow \infty} \frac{1}{n^9} = 0 \quad (6.6.11)$$

where we used the standard null sequence (i) with $p = 9$. It then follows from the properties of sequences tending to infinity that $a_n \rightarrow \infty$.

Alternatively, we could also note that:

$$a_n \geq \frac{2^n}{n} \quad (6.6.12)$$

but $\frac{2^n}{n} \rightarrow \infty$ so $a_n \rightarrow \infty$ by the squeeze rule.



Definition (Subsequence) The sequence (a_{n_k}) is a subsequence of the sequence (a_n) if (n_k) is a strictly increasing sequence of positive integers.

Example. Consider the sequence $a_n = n^{(-1)^n}$ then the odd subsequence is given by the terms:

$$a_{2n+1} = \frac{1}{2n+1} \quad (6.6.13)$$

whereas the even subsequence is given by:

$$a_{2n} = 2n \quad (6.6.14)$$



Theorem (Subsequence divergence)

For any subsequence (a_{n_k}) of a sequence (a_n) :

- (i) if $\lim_{n \rightarrow \infty} a_n = l$ then $\lim_{n \rightarrow \infty} a_{n_k} = l$.
- (ii) if $a_n \rightarrow \infty$ then $a_{n_k} \rightarrow \infty$

Proof.

- (i) Let $\epsilon > 0$, then there exists N such that:

$$|a_n - l| < \epsilon, \quad \forall n > N \quad (6.6.15)$$

Taking K such that $n_K \geq N$ then:

$$n_k \geq n_K \geq N, \quad \forall k > K \quad (6.6.16)$$

so that:

$$|a_{n_k} - l| < \epsilon, \quad \forall n_k > N \quad (6.6.17)$$

implying that $\lim_{n \rightarrow \infty} a_{n_k} = \lim_{n \rightarrow \infty} a_n = l$ as desired.

- (ii) Let $M > 0$, then there exists N such that:

$$a_n > M, \quad \forall n > N \quad (6.6.18)$$

Taking K such that $n_K \geq N$ then:

$$n_k \geq n_K \geq N, \quad \forall k > K \quad (6.6.19)$$

so that:

$$a_{n_k} > M, \quad \forall n_k > N \quad (6.6.20)$$

implying that $a_{n_k} \rightarrow \infty$ as desired. ■

Proposition (Subsequence rules)

The sequence (a_n) is divergent if (a_n) has two convergent subsequences with different limits.

The sequence (a_n) is divergent if at least one of its subsequences tends to infinity or minus infinity.

The above proposition follows immediately from writing the converse of the subsequence divergence theorem.

Example. Consider the sequence $a_n = \frac{n}{3} - \lfloor \frac{n}{3} \rfloor$ for $n = 1, 2, \dots$. Then the subsequence:

$$a_{3n} = n - \lfloor n \rfloor = 0 \quad (6.6.21)$$

so $\lim_{n \rightarrow \infty} a_{3n} = 0$. Instead:

$$a_{3n+1} = n + \frac{1}{3} - n = \frac{1}{3} \quad (6.6.22)$$

so $\lim_{n \rightarrow \infty} a_{3n+1} = \frac{1}{3}$. Since a_n has two subsequences converging to different values, we can conclude that a_n is divergent. \blacktriangleleft

Theorem (Convergent subsequence theorem) Let (a_n) be made up of two subsequences (a_{m_k}) and (a_{n_k}) which tend to the same limit l . Then:

$$\lim_{n \rightarrow \infty} a_n = l \quad (6.6.23)$$

Proof. Let $\epsilon > 0$, then there exists K_1 and K_2 such that:

$$|a_{m_k} - l| < \epsilon, \forall k > K_1 \quad (6.6.24)$$

$$|a_{n_k} - l| < \epsilon, \forall k > K_2 \quad (6.6.25)$$

If we let $N = \max\{K_1, K_2\}$ then:

$$|a_n - l| < \epsilon, \forall k > N \quad (6.6.26)$$

since for all $n > N$, $a_n = a_{m_k}$ or $a_n = a_{n_k}$, in which case the inequality is satisfied since $N \geq K_1$ and $N \geq K_2$. \blacksquare

6.7 Monotone convergence theorem

Theorem (Monotone convergence theorem)

If the sequence (a_n) is either:

(i) increasing and bounded above

(ii) decreasing and bounded below

then (a_n) is convergent.

Proof. Suppose (a_n) is bounded above, so that $\max\{a_n : n = 1, 2, \dots\} = l$, and let $\epsilon > 0$. Then there exists an integer N such that

$$a_N > l - \epsilon, \forall n > N \quad (6.7.1)$$

since otherwise $l - \epsilon$ would be an upper bound of a_n . Since a_n is increasing, $a_n \geq a_N$ for $n > N$ so that:

$$a_n > l - \epsilon \iff l - a_n < \epsilon, \forall n > N \quad (6.7.2)$$

Hence:

$$|a_n - l| = l - a_n < \epsilon, \forall n > N \quad (6.7.3)$$

proving that (a_n) converges to l . ■

Moreover, note that if (a_n) is increasing but not bounded above, then $a_n \rightarrow \infty$. Indeed, if it did converge to some value, then for any $\epsilon > 0$ there exists N such that:

$$|a_n - l| < \epsilon, \forall n > N \quad (6.7.4)$$

but since (a_n) is increasing but not bounded above, we can use the triangle inequality to write that:

$$|a_n| < \epsilon + |l|, \forall n > N \quad (6.7.5)$$

proving boundedness for all terms after N . If we define $M = \max\{|a_1|, |a_2|, \dots, |a_N|, 1 + |l|\}$ then:

$$|a_n| \leq M \quad (6.7.6)$$

which is a contradiction, since it was assumed a_n is unbounded.

We may restate this theorem as follows:

Theorem (Monotonic sequence theorem) A monotonic sequence is either convergent or diverges to ∞ .

Series

7.1 Introduction to Series

Definition (Series)

Let (a_n) be a sequence. Then the expression:

$$\sum_{i=m}^{\infty} a_i \quad (7.1.1)$$

an infinite series, and define by n th partial sum:

$$s_n = \sum_{i=m}^n a_i \quad (7.1.2)$$

The series is said to be convergent to s if its sequence converges to s , that is if:

$$\lim_{n \rightarrow \infty} \left(\sum_{i=m}^n a_i \right) = s \quad (7.1.3)$$

Otherwise it is said to be divergent.

Proposition (Geometric series)

The series of the form:

$$\sum_{k=0}^{\infty} ar^n \quad (7.1.4)$$

for some constants $a, r \in \mathbb{R}$ are called geometric series, and converge to:

$$\sum_{k=0}^{\infty} ar^n = \frac{a}{1-r} \text{ if } |r| < 1 \quad (7.1.5)$$

If $a \neq 0$ and $|r| \geq 1$, then the geometric series diverges.

Proof. **Lemma.** For $r \neq 1$, the partial sums are given by:

$$s_n = \sum_{k=0}^n ar^k = a \frac{1 - r^{n+1}}{1 - r} \quad (7.1.6)$$

Indeed, note that if $r \neq 1$ then:

$$(1 - r) \sum_{k=0}^n ar^k = \sum_{k=0}^n ar^k - \sum_{k=0}^n ar^{k+1} \quad (7.1.7)$$

$$= (a + ar + ar^2 + \dots + ar^n) - (ar + ar^2 + \dots + ar^n + ar^{n+1}) \quad (7.1.8)$$

$$= a + ar^{n+1} \quad (7.1.9)$$

$$\Rightarrow \sum_{k=0}^n ar^k = a \frac{1 - r^{n+1}}{1 - r} \quad (7.1.10)$$

as desired.

Hence, for $|r| < 1$, $\lim_{n \rightarrow \infty} r^{n+1} = 0$ so that $\lim_{n \rightarrow \infty} s_n = \frac{a}{1-r}$. ■

Definition (Cauchy criterion)

A series $\sum a_n$ satisfies the Cauchy criterion if its partial sums form a Cauchy sequence:

$$\forall \epsilon > 0, \exists N \text{ such that } m, n > N \implies |s_n - s_m| < \epsilon \quad (7.1.11)$$

7.2 Telescoping series

Telescoping series are series with terms of the form:

$$a_n = b_n - b_{n+i} \quad (7.2.1)$$

for some natural number i . The partial sums of a_n are:

$$s_n = \sum_{k=1}^n (b_k - b_{n \pm i}) \quad (7.2.2)$$

$$= (b_1 + \dots + b_{i+1} + b_{i+2} + \dots + b_n) - (b_{i+1} + b_{i+2} + \dots + b_{i+n}) \quad (7.2.3)$$

$$= (b_1 + b_2 + \dots + b_i + b_n + b_{n+1} + \dots + b_{n+i}) \quad (7.2.4)$$

Telescoping series most naturally occur with rational series. We provide an example below.

Example. Consider the series:

$$\sum_{n=1}^{\infty} \frac{1}{2n(2n+2)} \quad (7.2.5)$$

We can expand the partial fraction as:

$$\frac{1}{2n(2n+2)} = \frac{1}{2} \left(\frac{1}{2n} - \frac{1}{2n+2} \right) \quad (7.2.6)$$

so that the partial sums turn into:

$$s_n = \frac{1}{2} \sum_{k=1}^n \left(\frac{1}{2k} - \frac{1}{2k+2} \right) \quad (7.2.7)$$

$$= \frac{1}{2} \left[\left(\frac{1}{2} - \frac{1}{4} \right) + \left(\frac{1}{4} - \frac{1}{6} \right) + \dots + \left(\frac{1}{2n-2} - \frac{1}{2n} \right) + \left(\frac{1}{2n} - \frac{1}{2n+2} \right) \right] \quad (7.2.8)$$

$$= \frac{1}{2} \left(\frac{1}{2} - \frac{1}{2n+2} \right) = \frac{n}{4(n+1)} \quad (7.2.9)$$

We quickly see that the series does indeed converge:

$$\lim_{n \rightarrow \infty} s_n = \frac{1}{4} \quad (7.2.10)$$



7.3 Manipulating series

Proposition (Linearity of convergent series)

Suppose that $\sum_{n=1}^{\infty} a_n = s$ and $\sum_{n=1}^{\infty} b_n = t$ then:

$$\sum_{n=1}^{\infty} (\lambda a_n + b_n) = \lambda s + t \quad \forall \lambda \in \mathbb{R} \quad (7.3.1)$$

Proof. The partial sums associated to the two series are:

$$s_n = \sum_{k=1}^n a_k \quad \text{and} \quad t_n = \sum_{k=1}^n b_k \quad (7.3.2)$$

Then:

$$\sum_{k=1}^{\infty} (\lambda a_k + b_k) = (\lambda a_1 + b_1) + (\lambda a_2 + b_2) + \dots + (\lambda a_n + b_n) \quad (7.3.3)$$

$$= \lambda(a_1 + a_2 + \dots + a_n) + (b_1 + b_2 + \dots + b_n) \quad (7.3.4)$$

$$= \lambda s_n + t_n \quad (7.3.5)$$

By theorem 6.3:

$$\lim_{n \rightarrow \infty} (\lambda s_n + t_n) = \lambda \lim_{n \rightarrow \infty} s_n + \lim_{n \rightarrow \infty} t_n = \lambda s + t \quad (7.3.6)$$

as desired, the series is convergent to:

$$\sum_{n=1}^{\infty} (\lambda a_n + b_n) = \lambda s + t \quad (7.3.7)$$

■

Example. Consider the series:

$$\sum_{n=1}^{\infty} \left(\left(-\frac{3}{4} \right)^n - \frac{2}{n(n+1)} \right) = \sum_{n=1}^{\infty} \left(-\frac{3}{4} \right)^n - 2 \sum_{n=1}^{\infty} \frac{1}{n(n+1)} \quad (7.3.8)$$

The first is a convergent geometric series since $|r| = \frac{3}{4} < 1$. It converges to:

$$\sum_{n=1}^{\infty} \left(-\frac{3}{4} \right)^n = \frac{-\frac{3}{4}}{1 + \frac{3}{4}} = -\frac{3}{7} \quad (7.3.9)$$

Instead, the second series is a telescoping series whose partial sums can be expanded as:

$$s_n = \sum_{n=1}^n \frac{1}{n(n+1)} = \sum_{n=1}^n \left(\frac{1}{n} - \frac{1}{n+1} \right) \quad (7.3.10)$$

$$= \left(1 - \frac{1}{2} \right) + \left(\frac{1}{2} - \frac{1}{3} \right) + \dots + \left(\frac{1}{n-1} - \frac{1}{n} \right) + \left(\frac{1}{n} - \frac{1}{n+1} \right) \quad (7.3.11)$$

$$= 1 - \frac{1}{n+1} = \frac{n}{n+1} \quad (7.3.12)$$

so that:

$$\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = \lim_{n \rightarrow \infty} s_n = 1 \quad (7.3.13)$$

Hence the original series converges to:

$$\sum_{n=1}^{\infty} \left(\left(-\frac{3}{4} \right)^n - \frac{2}{n(n+1)} \right) = -\frac{3}{7} - 2 = -\frac{17}{7} \quad (7.3.14)$$

Theorem (Non-null test)

If $\sum_{n=1}^{\infty} a_n$ is convergent, then (a_n) is a null sequence.

If (a_n) is not a null sequence, then $\sum_{n=1}^{\infty} a_n$ is divergent.

Proof. Note that the second line is true provided the first line is true.

Let s_n be the n th partial sum of a_n . Since $\sum a_n$ converges, s_n must also converge to some limit s . Note that:

$$a_n = s_n - s_{n-1} \implies \lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} s_n - \lim_{n \rightarrow \infty} s_{n-1} = s - s = 0 \quad (7.3.15)$$

so that a_n is indeed a null sequence. ■

Example. Consider the series:

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1} n^2}{2n^2 + 1} \quad (7.3.16)$$

and let us examine the sequence (a_n) given by:

$$a_n = \frac{(-1)^{n+1} n^2}{2n^2 + 1} \quad (7.3.17)$$

Clearly, we can define a subsequence a_{2n} as consisting of all even terms:

$$a_{2n} = -\frac{n^2}{2n^2 + 1} \quad (7.3.18)$$

which is convergent to:

$$\lim_{n \rightarrow \infty} a_{2n} = -\lim_{n \rightarrow \infty} \frac{1}{2 + \frac{1}{n^2}} = \frac{1}{2 + \lim_{n \rightarrow \infty} \frac{1}{n^2}} = \frac{1}{2} \quad (7.3.19)$$

Therefore a_n is not a null sequence since it has a non-null subsequence, and thus the sum $\sum_{n=1}^{\infty} \frac{(-1)^{n+1} n^2}{2n^2 + 1}$ diverges.

7.4 Non-negative series

We continue our study of series by examining those containing only positive terms.

Example. Consider the series

$$\sum_{n=1}^{\infty} \frac{1}{n} \quad (7.4.1)$$

known as the harmonic series. Let us write the first few terms as:

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \dots \quad (7.4.2)$$

$$= 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4} \right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} \right) + \dots \quad (7.4.3)$$

Let (s_n) be the sequence of partial sums of the harmonic series, and consider the subsequence (s_{2^n})

$$s_2 = 1 + \frac{1}{2} \quad (7.4.4)$$

$$s_4 = 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4} \right) \geq 1 + \frac{1}{2} \quad (7.4.5)$$

$$s_8 = 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4} \right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} \right) \quad (7.4.6)$$

$$s_{2^k} = \sum_{n=1}^{2^k} \frac{1}{n} \geq 1 + \frac{1}{2} + \frac{1}{2} + \dots + \frac{1}{2} = 1 + \frac{1}{2}k \quad (7.4.7)$$

Since $1 + \frac{1}{2}k \rightarrow \infty$ as $k \rightarrow \infty$ it follows from the Squeeze rule that s_{2^k} is a divergent sequence, and therefore non-null. The harmonic series therefore diverges.

◀

Example. Consider instead the series:

$$\sum_{n=1}^{\infty} \frac{1}{n^2} \quad (7.4.8)$$

Consider the term:

$$\frac{1}{k^2} < \frac{1}{k(k-1)} = \frac{1}{k-1} - \frac{1}{k} \quad k > 1 \quad (7.4.9)$$

Therefore the k th partial sum of the series is:

$$s_n = \sum_{k=1}^n \frac{1}{k^2} < 1 + \sum_{k=2}^n \left(\frac{1}{k-1} - \frac{1}{k} \right) = 2 - \frac{1}{n} < 2 \quad (7.4.10)$$

Therefore, (s_n) is an increasing monotonic series that is bounded above. By the monotone convergence theorem it converges, so that:

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \lim_{n \rightarrow \infty} s_n \text{ converges} \quad (7.4.11)$$


Theorem (Comparison test)

If $0 \leq a_n \leq b_n$ for $n \in \mathbb{N}$ and $\sum b_n$ converges, then $\sum a_n$ converges too.

If instead $\sum a_n$ diverges then $\sum b_n$ diverges too.

Example. Consider the series:

$$\sum_{n=1}^{\infty} \frac{1}{n^3} \quad (7.4.12)$$

Then, we may write that:

$$0 \leq n^2 \leq n^3 \quad (7.4.13)$$

so that

$$0 \leq \frac{1}{n^3} \leq \frac{1}{n^2} \quad (7.4.14)$$

Since $\sum_{n=1}^{\infty} \frac{1}{n^2}$ is convergent, it follows from the Comparison test that $\sum_{n=1}^{\infty} \frac{1}{n^3}$ also converges.



Example. Consider the series:

$$\sum_{n=1}^{\infty} \frac{\cos^2 2n}{n^3} \quad (7.4.15)$$

Since $0 \leq \cos^2(2n) \leq 1$ we find that:

$$0 \leq \frac{\cos^2 2n}{n^3} \leq \frac{1}{n^3} \quad (7.4.16)$$

Since $\sum_{n=1}^{\infty} \frac{1}{n^3}$ converges, it follows from the Comparison test that $\sum_{n=1}^{\infty} \frac{\cos^2 2n}{n^3}$ also converges.


Theorem (Limit comparison test)

Suppose $\sum_{n=1}^{\infty} b_n$ converges. Suppose that $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} a_n$ are positive term series, such that

$$\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = L \neq 0 \quad (7.4.17)$$

If $\sum_{n=1}^{\infty} b_n$ is convergent, then $\sum_{n=1}^{\infty} a_n$ is convergent.

If $\sum_{n=1}^{\infty} b_n$ is divergent, then $\sum_{n=1}^{\infty} a_n$ is divergent.

Proof. Since $\frac{a_n}{b_n}$ is convergent, it must be bounded:

$$\frac{a_n}{b_n} \leq K \implies a_n \leq Kb_n, \quad n = 1, 2, \dots \quad (7.4.18)$$

Since $\sum_{n=1}^{\infty} b_n$ converges, it follows that $\sum_{n=1}^{\infty} Kb_n$ also converges, by the linearity of series. Hence, by the comparison test, $\sum_{n=1}^{\infty} a_n$ converges.

If instead $\sum_{n=1}^{\infty} b_n$ diverges, then note that:

$$\lim_{n \rightarrow \infty} \frac{b_n}{a_n} = \frac{1}{L} \neq 0 \quad (7.4.19)$$

since $L \neq 0$. Then:

$$\frac{b_n}{a_n} \leq K \implies b_n \leq Ka_n, \quad n = 1, 2, \dots \quad (7.4.20)$$

If $\sum_{n=1}^{\infty} a_n$ converges, it follows that $\sum_{n=1}^{\infty} Ka_n$ also converges, by the linearity of series. Hence:

$$\sum_{n=1}^{\infty} a_n \text{ converges} \implies \sum_{n=1}^{\infty} b_n \text{ converges} \quad (7.4.21)$$

which is equivalent to its converse:

$$\sum_{n=1}^{\infty} b_n \text{ diverges} \implies \sum_{n=1}^{\infty} a_n \text{ diverges} \quad (7.4.22)$$

as desired. ■

Example. Consider the series:

$$\sum_{n=1}^{\infty} \frac{1}{n^3 + n} = \sum_{n=1}^{\infty} a_n \quad (7.4.23)$$

Then, define:

$$b_n = \frac{1}{n^3} \quad (7.4.24)$$

so that:

$$\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = \lim_{n \rightarrow \infty} \frac{n^3}{n^3 + n} = \lim_{n \rightarrow \infty} \frac{1}{1 + \frac{1}{n^2}} = 1 \neq 0 \quad (7.4.25)$$

Therefore, by the Limit comparison test, it follows that since $\sum_{n=1}^{\infty} b_n$ converges as was shown earlier, $\sum_{n=1}^{\infty} \frac{1}{n^3+n}$ must also converge.



Example. Consider the series:

$$\sum_{n=1}^{\infty} \frac{n+4}{2n^3 - n + 1} = \sum_{n=1}^{\infty} a_n \quad (7.4.26)$$

Then we may define

$$b_n = \frac{1}{n^2} \quad (7.4.27)$$

so that:

$$\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = \lim_{n \rightarrow \infty} \frac{n^3 + 4n^2}{2n^3 - n + 1} = \lim_{n \rightarrow \infty} \frac{1 + \frac{4}{n}}{2 - \frac{1}{n^3} + \frac{1}{n^2}} = \frac{1}{2} \neq 0 \quad (7.4.28)$$

However, it was shown that $\sum_{n=1}^{\infty} \frac{1}{n^2}$ is convergent, so that $\sum_{n=1}^{\infty} \frac{n+4}{2n^3 - n + 1}$ also converges. ◀

Theorem (Ratio test) Suppose $\sum_{n=1}^{\infty} a_n$ is a series with positive terms. Then:

- (i) if $\lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n} = l$ with $0 \leq l < 1$ then $\sum_{n=1}^{\infty} a_n$ converges
- (ii) if $\lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n} = l$ with $l > 1$ then $\sum_{n=1}^{\infty} a_n$ diverges
- (iii) if $\lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n} = \infty$ then $\sum_{n=1}^{\infty} a_n$ diverges.

Proof.

- (i) Since $0 \leq l < 1$ we can choose $\epsilon > 0$ such that

$$l + \epsilon < 1 \quad (7.4.29)$$

If we let $r = l + \epsilon$, then since $r > l$ there exists N such that:

$$\frac{a_{n+1}}{a_n} \leq r, \quad \forall n \geq N \quad (7.4.30)$$

Then:

$$\frac{a_n}{a_N} = \prod_{k=n-1}^N \frac{a_{k+1}}{a_k} \leq \prod_{k=n-1}^N r = r^{n-N} \quad (7.4.31)$$

Hence:

$$a_n \leq a_N r^{n-N} \quad (7.4.32)$$

Note however that:

$$\sum_{n=1}^{\infty} a_N r^{n-N} \quad (7.4.33)$$

is a geometric series, and therefore converges. From the comparison test, it follows that $\sum_{n=1}^{\infty} a_n$ also converges.

(ii) and (iii) Suppose that:

$$\frac{a_{n+1}}{a_n} \rightarrow \infty \text{ or } \frac{a_{n+1}}{a_n} \rightarrow l \quad (7.4.34)$$

with $l > 1$ then there exists N such that:

$$\frac{a_{n+1}}{a_n} \geq 1, \forall n \geq N \quad (7.4.35)$$

Therefore:

$$\frac{a_n}{a_N} = \prod_{k=n-1}^N \frac{a_{k+1}}{a_k} \leq 1 \quad (7.4.36)$$

implying that:

$$a_n \geq a_N > 0, \forall n \geq N \quad (7.4.37)$$

(a_n) therefore cannot be a null sequence, and hence $\sum_{n=1}^{\infty} a_n$ must diverge. ■

Example. Consider the series:

$$\sum_{n=1}^{\infty} \frac{(2n)!}{n^n} = \sum_{n=1}^{\infty} a_n \quad (7.4.38)$$

Then:

$$\frac{a_{n+1}}{a_n} = \frac{(2n+2)!}{(n+1)^{n+1}} \cdot \frac{n^n}{(2n)!} \quad (7.4.39)$$

$$= (2n+2)(2n+1) \frac{n^n}{(n+1)^{n+1}} \quad (7.4.40)$$

$$= \frac{(2n+2)(2n+1)}{n+1} \frac{1}{(1+\frac{1}{n})^n} \quad (7.4.41)$$

Therefore:

$$\lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n} = \frac{2}{e} \lim_{n \rightarrow \infty} (2n+1) = \infty \quad (7.4.42)$$

so $\sum_{n=1}^{\infty} \frac{(2n)!}{n^n}$ diverges. ◀

Proposition (Standard series)

The following series converge:

- (i) $\sum_{n=1}^{\infty} \frac{1}{n^p}$ for $p \geq 2$
- (ii) $\sum_{n=1}^{\infty} c^n$ for $0 \leq c < 1$
- (iii) $\sum_{n=1}^{\infty} n^p c^n$ for $p > 0$ and $0 \leq c < 1$
- (iv) $\sum_{n=1}^{\infty} \frac{c^n}{n!}$ for $c \geq 0$

The following series is divergent:

- (v) $\sum_{n=1}^{\infty} \frac{1}{n^p}$ for $0 < p \leq 1$

Proof.

- (i) Note that if $p \geq 2$ then:

$$\frac{1}{n^p} \leq \frac{1}{n^2}, \quad n = 1, 2, \dots \quad (7.4.43)$$

and since $\sum_{n=1}^{\infty} \frac{1}{n^2}$ converges, by the Comparison test the series $\sum_{n=1}^{\infty} \frac{1}{n^p}$ must also converge.

- (ii) This is the standard geometric series with common ratio $r = c$, which converges provided $0 \leq c < 1$.

- (iii) Let $\sqrt{c} = b$, then:

$$a_n = (n^p b^n) b^n, \quad n = 1, 2, \dots \quad (7.4.44)$$

Since $0 \leq b < 1$, $n^p b^n$ is a standard null sequence. Setting $\epsilon = 1$, there exists N such that we have:

$$n^p b^n < 1, \quad \forall n > N \quad (7.4.45)$$

and thus:

$$a_b < b^n, \quad \forall n > N \quad (7.4.46)$$

However, $\sum_{n=1}^{\infty} b_n$ is a convergent geometric series, so by the Comparison test $\sum_{n=1}^{\infty} a_n$ converges.

- (iv) If $c = 0$, then convergence is trivial. Suppose $c \neq 0$, then:

$$\frac{a_{n+1}}{a_n} = \frac{c^{n+1}}{(n+1)!} \cdot \frac{c^n}{n!} = \frac{c}{n+1} \quad (7.4.47)$$

which as $n \rightarrow 0$ converges to 0. We then deduce from the ratio test that $\sum_{n=1}^{\infty} \frac{c^n}{n!}$ converges.

- (v) Note that if $p \leq 1$ then:

$$\frac{1}{n^p} \geq \frac{1}{n^2}, \quad n = 1, 2, \dots \quad (7.4.48)$$

and since $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges, by the Comparison test the series $\sum_{n=1}^{\infty} \frac{1}{n^p}$ must also diverge. ■

7.5 Series with positive and negative terms

Definition (*Absolute convergence*)

The series $\sum_{n=1}^{\infty} a_n$ is **absolutely convergent** if $\sum_{n=1}^{\infty} |a_n|$ is convergent.

Theorem (*Absolute convergence test*)

If $\sum_{n=1}^{\infty} a_n$ is absolutely convergent, then $\sum_{n=1}^{\infty} a_n$ is convergent.

Proof. Suppose that $\sum_{n=1}^{\infty} |a_n|$ converges, and let us define two new series $\sum_{n=1}^{\infty} b_n$ and $\sum_{n=1}^{\infty} c_n$ such that:

$$b_n = \begin{cases} a_n, & \text{if } a_n \geq 0 \\ 0, & \text{if } a_n < 0 \end{cases}, \quad \text{and } c_n = \begin{cases} 0, & \text{if } a_n \geq 0 \\ -a_n, & \text{if } a_n < 0 \end{cases} \quad (7.5.1)$$

then both b_n and c_n are non-negative. Moreover, $b_n \leq |a_n|$ and $c_n \leq |a_n|$ for $n = 1, 2, \dots$, so by the Comparison theorem, since $\sum_{n=1}^{\infty} |a_n|$ converges, it follows that $\sum_{n=1}^{\infty} b_n$ and $\sum_{n=1}^{\infty} c_n$ converge. Thus:

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} (b_n - c_n) = \sum_{n=1}^{\infty} b_n - \sum_{n=1}^{\infty} c_n \quad (7.5.2)$$

also converges. ■

Example. Consider the series:

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1} n}{n^3 + 1} \quad (7.5.3)$$

and its absolute analogue:

$$\sum_{n=1}^{\infty} \frac{n}{n^3 + 1} \quad (7.5.4)$$

Now note that if we define $a_n = \frac{n}{n^3 + 1}$ and $b_n = \frac{1}{n^2}$ then:

$$\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = \lim_{n \rightarrow \infty} \frac{n^3}{n^3 + 1} = 1 \neq 0 \quad (7.5.5)$$

By the limit comparison theorem, since $\sum_{n=1}^{\infty} \frac{1}{n^2}$ converges, it follows that $\sum_{n=1}^{\infty} \frac{n}{n^3+1}$ converges. Hence by the absolute convergence test $\sum_{n=1}^{\infty} \frac{(-1)^{n+1} n}{n^3+1}$ must also converge. \blacktriangleleft

Proposition (Triangle inequality) If $\sum_{n=1}^{\infty} a_n$ is absolutely convergent:

$$\left| \sum_{n=1}^{\infty} a_n \right| \leq \sum_{n=1}^{\infty} |a_n| \quad (7.5.6)$$

Proof. Suppose that $\sum_{n=1}^{\infty} |a_n|$ converges, and let us define two new series $\sum_{n=1}^{\infty} b_n$ and $\sum_{n=1}^{\infty} c_n$ such that:

$$b_n = \begin{cases} a_n, & \text{if } a_n \geq 0 \\ 0, & \text{if } a_n < 0 \end{cases}, \quad \text{and } c_n = \begin{cases} 0, & \text{if } a_n \geq 0 \\ -a_n, & \text{if } a_n < 0 \end{cases} \quad (7.5.7)$$

then both b_n and c_n are non-negative. Then, since:

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} b_n - \sum_{n=1}^{\infty} c_n \quad (7.5.8)$$

we find that

$$-\sum_{n=1}^{\infty} c_n \leq \sum_{n=1}^{\infty} a_n \leq \sum_{n=1}^{\infty} b_n \quad (7.5.9)$$

Thus, since $c_n \leq |a_n|$ and $b_n \leq |a_n|$ we find

$$-\sum_{n=1}^{\infty} |a_n| \leq \sum_{n=1}^{\infty} a_n \leq \sum_{n=1}^{\infty} |a_n| \implies \left| \sum_{n=1}^{\infty} a_n \right| \leq \sum_{n=1}^{\infty} |a_n| \quad (7.5.10)$$

as required. \blacksquare

Example. Using the absolute convergence test it is immediate that:

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{2^n} \quad (7.5.11)$$

converges. Note that:

$$\left| \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{2^n} \right| \leq \sum_{n=1}^{\infty} \frac{1}{2^n} = 1 \quad (7.5.12)$$

so it follows that the value at which the sum converges must lie in the interval $[-1, 1]$. \blacktriangleleft

Theorem (Alternating series test)

Let $a_n = (-1)^{n+1}b_n$ for $n = 1, 2, \dots$ then if (b_n) is a decreasing null sequence with positive terms:

$$\sum_{n=1}^{\infty} a_n \text{ converges} \quad (7.5.13)$$

Proof. Let us write the partial sum s_{2k} of the series as:

$$s_{2k} = (b_1 - b_2) + (b_3 - b_4) + \dots + (b_{2k-1} + b_{2k}) \quad (7.5.14)$$

Since (b_n) is decreasing, each bracket evaluates to a non-negative value. Consequently (s_{2n}) is an increasing sequence. Moreover:

$$s_{2k} = b_1 - (b_2 - b_3) - (b_4 - b_5) - \dots - (b_{2k-2} - b_{2k-1}) - b_{2k} \leq b_1 \quad (7.5.15)$$

By the monotone convergence theorem it follows that

$$\lim_{n \rightarrow \infty} s_{2n} = s \quad (7.5.16)$$

for some s . Also:

$$s_{2k-1} = b_1 - (b_2 - b_3) - (b_4 - b_5) - \dots - (b_{2k-2} - b_{2k-1}) = s_{2k} + b_{2k} \quad (7.5.17)$$

so that:

$$\lim_{n \rightarrow \infty} s_{2k+1} = \lim_{n \rightarrow \infty} s_{2k} + \lim_{n \rightarrow \infty} b_{2k} = s \quad (7.5.18)$$

where since b_n is null, all its subsequences are null. Since both the odd and even subsequences converge to s , we find that $s_n \rightarrow s \implies \sum_{n=1}^{\infty} a_n = s$. ■

Example. Consider the sequence:

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n + \sqrt{n}} \quad (7.5.19)$$

We can write its terms as:

$$a_n = \frac{(-1)^{n+1}}{n + \sqrt{n}} = (-1)^{n+1} \frac{1}{n + \sqrt{n}} = (-1)^{n+1} b_n \quad (7.5.20)$$

Now:

- (i) $b_n = \frac{1}{n + \sqrt{n}} \geq 0$ for $n = 1, 2, \dots$
- (ii) Since:

$$\frac{1}{n + \sqrt{n}} \leq \frac{1}{n} \quad (7.5.21)$$

and $\frac{1}{n}$ is a null sequence, by the Squeeze theorem (b_n) is a null sequence.

(iii) (b_n) is decreasing, since $\left(\frac{1}{b_n}\right) = n + \sqrt{n}$ is increasing. Hence, by the alternating test:

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n + \sqrt{n}} \quad (7.5.22)$$

converges.



Unit D4: Functions and Continuity

8.1 Real functions

Definition (*Real function*)

Let $A \subseteq \mathbb{R}$ and $f : A \rightarrow \mathbb{R}$ is a bijective function. Then the inverse function $f^{-1} : \text{Im}(A) \rightarrow A$ and has rule:

$$f^{-1}(f(x)) = x, \quad \forall x \in A \quad (8.1.1)$$

Example. Consider the function:

$$f(x) = \frac{1}{1-x}, \quad \forall x \in (-\infty, 1) \quad (8.1.2)$$

Let's solve the equation $y = f(x)$.

$$y = \frac{1}{1-x} \iff \frac{1}{y} = 1-x \iff x = 1 - \frac{1}{y} \quad (8.1.3)$$

so every value of $y \in \text{Im}(f)$ is the image of one $x \in A$, showing that f is bijective, and thus invertible. Its inverse is clearly:

$$f^{-1}(x) = 1 - \frac{1}{x} \quad (8.1.4)$$

To determine the domain, we firstly find the image of f . Since $x < -1$, we must have that:

$$\text{Im}(f) = f(A) = \left\{ \frac{1}{1-x} : \forall x < -1 \right\} = (0, \infty) \quad (8.1.5)$$

so the domain of f^{-1} must be $(0, \infty)$.



Unfortunately, often times it is hard to solve $y = f(x)$ for x . In such instances, one can use the following theorem to prove that a function is bijective.

Theorem (Invertibility of monotonic functions)

If a function f is strictly increasing or strictly decreasing on some interval A , then it is invertible on A .

Example. Consider the function:

$$f(x) = x^5 + x - 1, \quad \forall x \in \mathbb{R} \quad (8.1.6)$$

If $x_1 < x_2$, then $x_1^5 < x_2^5$ so:

$$x_1^5 + x_1 < x_2^5 + x_2 \implies x_1^5 + x_1 - 1 < x_2^5 + x_2 \implies f(x_1) < f(x_2) \quad (8.1.7)$$

Thus, $f(x)$ is strictly increasing, and thus invertible. ◀

8.2 Continuity

Consider some real function f , one important question to ask about this function is whether or not it has any weird gaps, jumps, whether its graph can be drawn without lifting the pen from the paper.

Intuitively, we can define continuity as a property of a function f such that if (x_n) is any sequence on its domain that tends to a , then $f(x_n)$ tends to $f(a)$.

In the case of a jump for example, $x_n \rightarrow a$, yet $f(x_n)$ has no limit, because the subsequence $f(x_n)$ with $x_n < a$ and the subsequence $f(x_n)$ with $x_n > a$ do not converge to the same value.

Definition (Continuity)

A function $f : A \rightarrow \mathbb{R}$ is continuous at $a \in A$ if $\forall (x_n) \in A$ such that $x_n \rightarrow a$, we have $f(x_n) \rightarrow f(a)$.

If f is not continuous at a , it is discontinuous at a .

We say that f is continuous on A if it is continuous for all points in A .

Example. Consider the function:

$$f(x) = x^3 - 2x^2 \quad (8.2.1)$$

at the point $a = 2$. Consider any sequence $x_n \rightarrow 2$, then:

$$f(x_n) = x_n^3 - 2x_n^2 \rightarrow 2^3 - 2 \cdot 2^2 = 0 \quad (8.2.2)$$

using the limit combination properties. Moreover, $f(2) = 0$ thus f is indeed continuous at $a = 2$. ◀

Example. Consider the function:

$$f(x) = \lfloor x \rfloor \quad (8.2.3)$$

at the point $a = 1$. Consider the sequence $x_n = 1 - \frac{1}{n}$, so that $x_n \rightarrow 1$. Then:

$$f(x_n) = \left\lfloor 1 - \frac{1}{n} \right\rfloor = 0 \quad (8.2.4)$$

since for any $n \geq 1$, $1 - \frac{1}{n} < 1$ and thus $\left\lfloor 1 - \frac{1}{n} \right\rfloor = 0$. However, $f(1) = 1 \neq 0$, thus the function is discontinuous at $a = 1$. \blacktriangleleft

Example. Consider the function:

$$f(x) = \begin{cases} \sin \frac{1}{x}, & x \neq 0 \\ 0, & x = 0 \end{cases} \quad (8.2.5)$$

Note that:

$$\sin \left(2n + \frac{1}{2} \right) \pi = 1 \quad (8.2.6)$$

so if we define $x_n = \frac{1}{(2n + \frac{1}{2})\pi}$ then:

$$\sin x_n \rightarrow 1 \neq 0 = f(0) \quad (8.2.7)$$

so we see that f is discontinuous at $x = 0$. \blacktriangleleft

Example. Consider the function:

$$f(x) = |x|, \forall x \in \mathbb{R} \quad (8.2.8)$$

Consider any $a \in \mathbb{R}$ and let (x_n) be any sequence in \mathbb{R} such that $x_n \rightarrow a$ as $n \rightarrow \infty$. Now using the triangle inequality:

$$|x_n - a| \geq ||x_n| - |a||, n = 1, 2, \dots \quad (8.2.9)$$

and since $x_n - a$ is a null sequence, we must have that $|x_n| - |a|$ is also a null sequence. Thus, $|x_n| \rightarrow |a| = f(a)$ as desired. Thus f is continuous everywhere on \mathbb{R} . \blacktriangleleft

Example. Consider the function:

$$f(x) = \sqrt{x}, \forall x \in [0, \infty) \quad (8.2.10)$$

Consider any $a \in [0, \infty)$ and let (x_n) be any sequence in \mathbb{R} such that $x_n \rightarrow a$ as $n \rightarrow \infty$. Now since $x_n - a$ is a null sequence, $|x_n - a|$ is also a null sequence, and by the power rule $\sqrt{|x_n - a|}$ must also be a null sequence.

Also note that in unit D1 we derived:

$$\sqrt{|x_n - a|} \geq |\sqrt{x_n} - \sqrt{a}| \quad (8.2.11)$$

and thus $\sqrt{x_n} - \sqrt{a}$ is also a null sequence. Thus, $\sqrt{x_n} \rightarrow \sqrt{a} = f(a)$ as desired. Thus f is continuous everywhere on \mathbb{R} . \blacktriangleleft

8.3 Properties of continuous functions

Proposition (Combination of continuous functions)

Suppose f, g are continuous functions at a , then:

- (i) $f + g$
- (ii) αf , $\forall \alpha \in \mathbb{R}$
- (iii) fg
- (iv) f/g is $g(a) \neq 0$
- (v) $f \circ g$

Proof. We prove (v). Suppose f is continuous at a and g is continuous at $f(a)$. If f has domain A and g has domain B , so that the domain of $g \circ f$ is:

$$C = \{x \in A : f(x) \in B\} \ni a \quad (8.3.1)$$

We know that $f(x_n) \rightarrow a$ for all sequences $(x_n) \in A$, implying that $(f(x_n)) \in B$. Moreover, we know that g is continuous at $f(a)$ so that $g(f(x_n)) \rightarrow g(f(a))$, as desired. \blacksquare

The following theorem results immediately

Theorem (Continuity of polynomials and their rationals)

The following are continuous:

- (i) any polynomial $p(x) = a_0 + a_1x + \dots + a_nx^n$
- (ii) any rational function $r(x) = \frac{p(x)}{q(x)}$ where p, q are polynomials (over $\mathbb{R} - \{x : q(x) = 0\}$).

Example. Let's prove that:

$$f(x) = \sqrt{x^2 + 2x + 2} - \frac{3x}{x^4 + 4}, \quad \forall x \in \mathbb{R} \quad (8.3.2)$$

is continuous. To do so, first note that $x^2 + 2x + 2$ is a polynomial, and thus continuous. Moreover, it is always positive, since it has no real roots. Therefore, if we define $h(x) = x^2 + 2x + 2$, h is continuous on \mathbb{R} , and $g(x) = \sqrt{x}$ is continuous on $[0, \infty)$. Thus, $g(h(x))$ is also continuous on \mathbb{R} .

Similarly, $e(x) = x^4 + 4$ is continuous everywhere on \mathbb{R} since it is a polynomial, and has non-zero values, since $x^4 = -4$ has no real roots. Therefore, if we define $d(x) = \frac{3x}{x^4+4}$ must also be continuous on \mathbb{R} , since $3x$ and $x^4 + 4$ are (non-zero) polynomials. Hence, $f(x) = g(h(x)) + d(x)$ is continuous on \mathbb{R} by the combination rules. \blacktriangleleft

Theorem (Squeeze rule)

Let f, g, h be defined on an open interval I and let $a \in I$. If:

- (i) $g(x) \leq f(x) \leq h(x)$ for $x \in I$
- (ii) $g(a) = f(a) = h(a)$
- (iii) g, h are continuous at a

then f is continuous at a .

Proof. Suppose that f, g, h satisfy these conditions. Since $x_n \rightarrow a$, there must exist N such that:

$$x_n \in (a - \epsilon, a + \epsilon) \subseteq I, \forall n > N \quad (8.3.3)$$

Hence by condition 1:

$$g(x_n) \leq f(x_n) \leq h(x_n) \quad (8.3.4)$$

Condition 2,3 imply that:

$$\lim_{n \rightarrow \infty} g(x_n) = \lim_{n \rightarrow \infty} h(x_n) = f(a) \quad (8.3.5)$$

so by the Squeeze rule of sequences, $\lim_{n \rightarrow \infty} f(x_n) = f(a)$, and f is continuous at a . \blacksquare

Example. Consider the function:

$$f(x) = \begin{cases} x^2 \cos \frac{1}{x^2}, & x \neq 0 \\ 0, & x = 0 \end{cases} \quad (8.3.6)$$

Then, we know that:

$$-1 \leq \cos \frac{1}{x^2} \leq 1, \quad x \neq 0 \quad (8.3.7)$$

so that:

$$-x^2 \leq x^2 \cos \frac{1}{x^2} \leq x^2, \quad x \neq 0 \quad (8.3.8)$$

Now, since $-x^2 = 0 \leq f(0) = 0 \leq 0 = x^2$, we may assert that:

$$g(x) \leq f(x) \leq h(x) \quad (8.3.9)$$

where $g(x) = -x^2$ and $h(x) = x^2$. Moreover, $g(0) = h(0) = f(0) = 0$, and since g, h are polynomials they are continuous. Thus by the squeeze rule f must be continuous at $x = 0$. \blacktriangleleft

Theorem (Glue rule)

Let f be defined on an open interval I and let $a \in I$. If h, g are functions satisfying:

- (i) $f(x) = g(x)$ for $x \in I, x < a$, and $f(x) = h(x)$ for $x \in I, x > a$
- (ii) $f(a) = g(a) = h(a)$
- (iii) g, h are continuous at a

then f is continuous at a .

Proof. Suppose f, g, h satisfy the above conditions. Then:

$$x_n \in (a - \epsilon, a + \epsilon) \subseteq I, \forall n > N \quad (8.3.10)$$

since $x_n \rightarrow a$. We define $(x_n)^\infty_N$, consists of two subsequences (x_{m_k}) and (x_{n_k}) satisfying:

$$x_{m_k} < a, \text{ and } x_{n_k} \geq a \quad (8.3.11)$$

The conditions give:

$$g(x_{m_k}) \rightarrow g(a) = f(a), \text{ and } h(x_{n_k}) \rightarrow h(a) = g(a) \quad (8.3.12)$$

so that:

$$f(x_{m_k}) \rightarrow f(a), \text{ and } f(x_{n_k}) \rightarrow f(a) \quad (8.3.13)$$

Therefore, $f(x_n)$ consists of two subsequences convergent to $f(a)$, and thus $f(x_n) \rightarrow f(a)$, as desired. \blacksquare

Example. Consider:

$$f(x) = \begin{cases} x^3 - 3x + 5, & x < 1 \\ \frac{2x+1}{3x-2}, & x \geq 1 \end{cases} \quad (8.3.14)$$

Let's define $g(x) = x^3 - 3x + 5$ and $h(x) = \frac{2x+1}{3x-2}$, then:

$$f(x) = g(x), \text{ for } x < 1 \quad (8.3.15)$$

and

$$f(x) = h(x), \text{ for } x > 1 \quad (8.3.16)$$

Moreover, $f(1) = 3 = g(1) = h(1)$, and g, h are both continuous at $x = 1$ since the first is a polynomial and the second is the ratio of two polynomials, with non-zero determinant. Hence, by the Glue theorem, we have that f is continuous. \blacktriangleleft

8.4 Trigonometric and exponential functions

We will now prove that the function \sin, \cos, \tan and \exp are continuous.

Proposition (Sine inequality)

We have that:

$$\sin x \leq x, \text{ for } 0 \leq x \leq \frac{\pi}{2} \quad (8.4.1)$$

Proof. Consider the function $f(x) = x - \sin x$. If $x = 0$ then $f(x) = 0$. Moreover, for $0 \leq x \leq \frac{\pi}{2}$ we have that:

$$f'(x) = 1 - \cos x \quad (8.4.2)$$

and since $0 \leq \cos x \leq 1$ over this interval, we have that:

$$0 \leq f'(x) \leq 1 \quad (8.4.3)$$

so $f(x)$ is increasing over $0 \leq x \leq \frac{\pi}{2}$, and since $f(0) = 0$, $f(x) \geq 0$ implying that $\sin x \leq x$ as desired. \blacksquare

An important consequence of this inequality is the following.

Corollary. $|\sin x| \leq |x|$

Proof. The sine inequality proves this result for $0 \leq x \leq \frac{\pi}{2}$. For $x > \frac{\pi}{2}$:

$$|\sin x| \leq 1 < \frac{\pi}{2} < x = |x| \quad (8.4.4)$$

Finally, if $x < 0$ then:

$$|\sin(x)| = |\sin(-x)| \leq |-x| = |x| \quad (8.4.5)$$

as desired. \blacksquare

Theorem (Continuity of trigonometric functions)

The trigonometric functions \sin, \cos, \tan are continuous.

Proof. We need to show that:

$$\sin x_n \rightarrow \sin a \quad (8.4.6)$$

for all null sequences $x_n - a$.

We can use the property:

$$\sin x - \sin a = 2 \cos\left(\frac{1}{2}(x+a)\right) 2 \sin\left(\frac{1}{2}(x-a)\right) \quad (8.4.7)$$

So:

$$|\sin x_n - \sin a| = 2 \left| \cos\left(\frac{1}{2}(x_n+a)\right) 2 \sin\left(\frac{1}{2}(x_n-a)\right) \right| \quad (8.4.8)$$

$$\leq 2 \left| \sin\left(\frac{1}{2}(x_n-a)\right) \right| \quad (8.4.9)$$

$$\leq |x_n - a| \quad (8.4.10)$$

Since $x_n - a$ is a null sequence, we must have that $\sin x_n \rightarrow \sin a$ as desired.

Since $\cos x = \sin(x + \frac{\pi}{2})$, we can use the continuity of composite functions to state that it too must be continuous. The same goes for $\tan x = \frac{\sin x}{\cos x}$. ■

Proposition (Exponential inequalities)

- (i) $e^x \geq 1 + x$ for $x \geq 0$
- (ii) $e^x \leq \frac{1}{1-x}$ for $0 \leq x < 1$.

Proof. These follow immediately from the power series representation of e^x (for $x \geq 0$) and $\frac{1}{1-x}$ (for $0 \leq x < 1$). Indeed:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots \leq 1 + x \quad (8.4.11)$$

for $x \geq 0$. Similarly:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots \leq 1 + x + x^2 + x^3 + \dots = \frac{1}{1-x} \quad (8.4.12)$$

for $0 \leq x < 1$. ■

Corollary. $1 + x \leq e^x \leq \frac{1}{1-x}$ for $|x| < 1$.

Proof. We have already proven the case for $0 \leq x < 1$. Consider now the case for $-1 < x < 0 \implies 0 < -x < 1$. Then:

$$1 - x \leq e^{-x} \leq \frac{1}{1+x} \quad (8.4.13)$$

Since all terms in the inequality are non-zero on $0 < -x < 1$, we can take the reciprocal:

$$1 + x \leq e^x \leq 1 - x \quad (8.4.14)$$

Thus, the inequality has been proven. ■

Theorem (Continuity of the exponential function) The exponential function \exp is continuous.

Proof. We need to prove that for all sequences $x_n \rightarrow a$ we have $e^{x_n} \rightarrow e^a$.

If $x_n - a$ is a null sequence, then there exists N such that $|x_n - a| \leq 1$ for $n > N$ (we use $\epsilon = 1$). Therefore:

$$1 + (x_n + a) \leq e^{x_n - a} \leq \frac{1}{1 - (x_n - a)}, \quad \forall n > N \quad (8.4.15)$$

By the squeeze rule, we find that $e^{x_n - a} \rightarrow 1$ and thus $e^{x_n} \rightarrow e^a$ as desired. ■

We summarize the main continuous functions we have found below:

Proposition (Standard continuous functions) The following are all continuous:

- (i) polynomials and rational functions of polynomials
- (ii) $f(x) = |x|$
- (iii) $f(x) = \sqrt{x}$
- (iv) the trigonometric functions
- (v) the exponential function

Unit F1: Limits

9.1 Introduction to limits of functions

Definition 9.1 (Punctured neighbourhood) The **punctured neighbourhood** of a point $c \in \mathbb{R}$ is a bounded open interval whose midpoint is c and has been removed. If we define the width of the neighborhood to be 2ε then:

$$N_\varepsilon(c) = (c - \varepsilon, c) \cup (c, c + \varepsilon) \quad (9.1.1)$$

The concept of a punctured neighborhood is essential in defining the limit of a function.

Definition 9.2 (Limit of a function)

Let f be a function defined on $N_\varepsilon(c)$. Then $f(x)$ tends to l as x tends to c if $l \in \mathbb{R}$ and for all sequences (x_n) in $N_\varepsilon(c)$ such that $x_n \rightarrow c$, we have that $f(x_n) \rightarrow l$. We write this as:

$$\lim_{x \rightarrow c} f(x) = l \quad (9.1.2)$$

Example. Let us try to prove that $\lim_{x \rightarrow 0} \frac{\sin x}{x} = 1$.

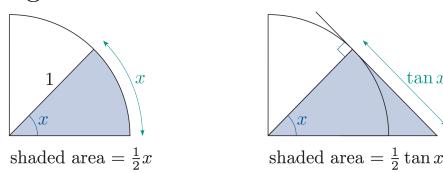
Firstly, in the previous unit we established the inequality:

$$\sin x \leq x, \quad \text{for } 0 < x \leq \frac{\pi}{2} \quad (9.1.3)$$

We may also deduce that:

$$x \leq \tan x, \quad \text{for } 0 < x \leq \frac{\pi}{2} \quad (9.1.4)$$

as can be seen from the figures below:



The first inequality (9.1.3) may be rearranged into the more useful form:

$$\frac{\sin x}{x} \leq 1, \quad \text{for } 0 < x \leq \frac{\pi}{2} \quad (9.1.5)$$

and similarly for the second inequality (9.1.4) may be written as:

$$\cos x \leq \frac{\sin x}{x}, \quad \text{for } 0 < x \leq \frac{\pi}{2} \quad (9.1.6)$$

Hence, these two may be combined into a single inequality providing upper and lower bounds for $\frac{\sin x}{x}$:

$$\cos x \leq \frac{\sin x}{x} \leq 1, \quad \text{for } 0 < x \leq \frac{\pi}{2} \quad (9.1.7)$$

Now since both $\cos x$ and $\frac{\sin x}{x}$ are both even functions, we may substitute $x' = -x$ into (9.1.7) to find that:

$$\cos x \leq \frac{\sin x}{x} \leq 1, \quad \text{for } 0 < |x| \leq \frac{\pi}{2} \quad (9.1.8)$$

Suppose (x_n) is a null sequence in the neighborhood $N_{\frac{\pi}{2}}(0)$ so that:

$$\cos x_n \leq \frac{\sin x_n}{x_n} \leq 1, \quad n = 1, 2, \dots \quad (9.1.9)$$

Using the squeeze rule, we see that since $\cos x_n \rightarrow 1$ it must be that $\frac{\sin x_n}{x_n} \rightarrow 1$. This proves that:

$$\lim_{x \rightarrow 0} \frac{\sin x}{x} = 1 \quad (9.1.10)$$

◀

Example. Consider $\lim_{x \rightarrow 0} |x|$, we will show that this limit does not exist. Indeed, consider the neighbourhood $N_{\frac{1}{2}}(1)$ and the sequences $x_n = 1 - \left(\frac{1}{2}\right)^n$ and $y_n = 1 + \left(\frac{1}{2}\right)^n$ defined on this neighbourhood. We see that:

$$\lim_{n \rightarrow \infty} f(x_n) = \lim_{n \rightarrow \infty} \left[1 - \left(\frac{1}{2}\right)^n \right] = 0 \quad (9.1.11)$$

whereas:

$$\lim_{n \rightarrow \infty} f(y_n) = \lim_{n \rightarrow \infty} \left[1 + \left(\frac{1}{2}\right)^n \right] = 1 \quad (9.1.12)$$

Therefore, we have found two different sequences on $N_{\frac{1}{2}}(0)$ which converge to different values, showing that $f(x) = |x|$ does not tend to a limit as $x \rightarrow 1$.

◀

Theorem 9.3 (Continuity \iff limit)

Let f be a function defined on an open interval I and let $c \in I$. Then:

$$f \text{ is continuous at } c \iff \lim_{x \rightarrow c} f(x) = f(c) \quad (9.1.13)$$

This is particularly useful when trying to determine the limit of a function that falls within the class of standard continuous functions, such as polynomials, exponential and logarithmic functions etc...

Proposition 9.4 (Composition rule)

If $\lim_{x \rightarrow c} f(x) = l$ and $\lim_{x \rightarrow l} g(x) = L$, then $\lim_{x \rightarrow c} g(f(x)) = L$ provided:

$$g \text{ is defined and continuous at } l \quad (9.1.14)$$

or

$$f(x) \neq l, \forall x \in N_\varepsilon(c) \quad (9.1.15)$$

Example. Let us try to evaluate $\lim_{x \rightarrow 0} \sqrt{\frac{x}{\sin x}}$. Let us define $f(x) = \frac{\sin x}{x}$ and $g(x) = \frac{1}{\sqrt{x}}$. Then, we have that:

$$\lim_{x \rightarrow 0} f(x) = 1 \quad (9.1.16)$$

as was found previously. Moreover

$$\lim_{x \rightarrow 1} g(x) = 1 \quad (9.1.17)$$

so that, since g is defined and continuous at 1 (using the composition rules of continuous functions):

$$\lim_{x \rightarrow 0} g(f(x)) = \lim_{x \rightarrow 0} \sqrt{\frac{x}{\sin x}} = 1 \quad (9.1.18)$$



Example. We consider the limit:

$$\lim_{x \rightarrow 0} \frac{1 - \cos x}{x} \quad (9.1.19)$$

We use the identity:

$$\cos x = 1 - 2 \sin^2 \frac{x}{2} \quad (9.1.20)$$

to find that:

$$\lim_{x \rightarrow 0} \frac{2 \sin^2 \frac{x}{2}}{x} = \lim_{x \rightarrow 0} \frac{\sin \frac{x}{2}}{\frac{x}{2}} \cdot \lim_{x \rightarrow 0} \sin \frac{x}{2} = 0 \quad (9.1.21)$$

where we used the substitution $u = \frac{x}{2}$ to evaluate $\lim_{x \rightarrow 0} \frac{\sin \frac{x}{2}}{\frac{x}{2}}$. ◀

Theorem 9.5 (Squeeze rule for limits)

Let f, g, h be functions defined on $N_\epsilon(c)$ for some $r > 0$. If:

- (i) $g(x) \leq f(x) \leq h(x), \forall x \in N_\epsilon(c)$
- (ii) $\lim_{x \rightarrow c} g(x) = \lim_{x \rightarrow c} h(x) = l$ then $\lim_{x \rightarrow c} f(x) = l$.

Example. We have shown previously that:

$$1 + x \leq e^x \leq \frac{1}{1 - x}, \quad \text{for } |x| < 1 \quad (9.1.22)$$

This may rearranged into:

$$1 + x \leq e^x \leq \frac{1}{1 - x}, \quad \text{for } |x| < 1 \quad (9.1.23)$$

$$1 \leq \frac{e^x - 1}{x} \leq \frac{e^x - 1}{x} \leq \frac{1}{1 - x}, \quad \text{for } 0 < |x| < 1 \quad (9.1.24)$$

Now we have that:

$$1 - \frac{|x|}{1 - x} = \frac{x - |x|}{1 - x} \leq 1 \quad (9.1.25)$$

since if $x < 0$ then $|x| - x = 2|x| > 0$ whereas if $x > 0$ then $|x| - x = 0$. Similarly:

$$\frac{1}{1 - x} \leq 1 + \frac{|x|}{1 - x} = \frac{1 + (|x| - x)}{1 - x} \quad (9.1.26)$$

Consequently, we have the following inequality:

$$1 - \frac{|x|}{1 - x} \leq \frac{e^x - 1}{x} \leq 1 + \frac{|x|}{1 - x} \quad (9.1.27)$$

Define $g(x) = 1 - \frac{|x|}{1 - x}$, $h(x) = 1 + \frac{|x|}{1 - x}$ and $f(x) = \frac{e^x - 1}{x}$ on $N_\epsilon(c)$ for some $\epsilon > 0$. Then the first condition of the squeeze rule is clearly satisfied:

$$g(x) \leq f(x) \leq h(x) \quad (9.1.28)$$

Next, we also have that:

$$\lim_{x \rightarrow 0} g(x) = 1 = \lim_{x \rightarrow 0} h(x) \quad (9.1.29)$$

Therefore we may conclude that

$$\lim_{x \rightarrow 0} \frac{e^x - 1}{x} = 1 \quad (9.1.30)$$

using the Squeeze rule. ◀

Definition 9.6 (One-sided limit)

Let $f(x)$ be a function defined on $(c, c+r)$ for $r > 0$. Then we say that $f(x)$ tends to l as x tends to c from the right:

$$\lim_{x \rightarrow c^+} f(x) = l \quad (9.1.31)$$

provided that for each sequence (x_n) in $(c, c+r)$ such that $x_n \rightarrow c$, $f(x_n) \rightarrow l$.

Let $f(x)$ be a function defined on $(c-r, c)$ for $r > 0$. Then we say that $f(x)$ tends to l as x tends to c from the left:

$$\lim_{x \rightarrow c^-} f(x) = l \quad (9.1.32)$$

provided that for each sequence (x_n) in $(c-r, c)$ such that $x_n \rightarrow c$, $f(x_n) \rightarrow l$.

The following result follows immediately from the fact that $(c-\epsilon, c) \cup (c, c+\epsilon) = N_\epsilon(c)$.

Theorem 9.6 (Ordinary limits and one sided-limits)

Let f be defined on $N_\epsilon(c)$ for $r > 0$. Then:

$$\lim_{x \rightarrow c} f(x) = l \iff \lim_{x \rightarrow c^+} f(x) = \lim_{x \rightarrow c^-} f(x) = l \quad (9.1.33)$$

Finally, we also have an analogue of Theorem 9.3 for one-sided limits:

Proposition 9.7 (Continuity \iff one-sided limit)

Let f be a function whose domain I is an interval with left-hand endpoint c included (so either $[c, \infty)$, $[c, b)$ or $[c, b]$ where $b > c$). Then:

$$f \text{ is continuous at } c \iff \lim_{x \rightarrow c^\pm} f(x) = f(c) \quad (9.1.34)$$

Example. Let us evaluate $\lim_{x \rightarrow 0^+} \left(\frac{\sin x}{x} + \sqrt{x} \right)$.

We have from typical combination rules that:

$$\lim_{x \rightarrow 0^+} \left(\frac{\sin x}{x} + \sqrt{x} \right) = \lim_{x \rightarrow 0^+} \frac{\sin x}{x} + \lim_{x \rightarrow 0^+} \sqrt{x} = 1 + 0 = 0 \quad (9.1.35)$$

where we used Theorem 9.6 to get:

$$\lim_{x \rightarrow 0} \frac{\sin x}{x} = 1 \implies \lim_{x \rightarrow 0^+} \frac{\sin x}{x} = 1 \quad (9.1.36)$$

and Proposition 9.7 to get

$$\lim_{x \rightarrow 0} \sqrt{x} = 1 \implies \lim_{x \rightarrow 0^+} \sqrt{x} = 1 \quad (9.1.37)$$



9.2 Asymptotic behaviour

Definition 9.8 (*Infinite limit*)

Let f be defined on $N_\epsilon(c)$. Then $f(x)$ tends to ∞ as x tends to c if for each sequence (x_n) in $N_\epsilon(c)$ such that $x_n \rightarrow c$, $f(x_n) \rightarrow \infty$. We write that:

$$f(x) \rightarrow \infty \text{ as } x \rightarrow c \quad (9.2.1)$$

Theorem 9.9 (*Reciprocal rule for limits*)

If f is a function satisfying:

(i) $f(x) > 0$ for $x \in N_\epsilon(c)$, where $\epsilon > 0$

(ii) $f(x) \rightarrow 0$ as $x \rightarrow c$

then $\frac{1}{f(x)} \rightarrow \infty$ as $x \rightarrow c$.

Example. Consider the asymptotic behaviour of $\frac{1}{x^3 - 1}$ as $x \rightarrow 1^+$. Define $f(x) = x^3 - 1$, then we have that:

$$f(x) = x^3 - 1 > 0 \quad \forall x \in (1, \infty) = (1, 1+r) \quad (9.2.2)$$

for some $r > 0$. Moreover

$$\lim_{x \rightarrow 1^+} f(x) = 0 \quad (9.2.3)$$

Therefore:

$$\frac{1}{x^3 - 1} \rightarrow \infty \text{ as } x \rightarrow 1^+ \quad (9.2.4)$$



Theorem 9.10 (*Squeeze rule for $x \rightarrow \infty$*)

Let f, g, h be defined on (R, ∞) . Then:

(a) if

(i) $g(x) \leq f(x) \leq h(x)$, $\forall x \in (R, \infty)$

(ii) $\lim_{x \rightarrow \infty} g(x) = \lim_{x \rightarrow \infty} h(x) = l$

then $\lim_{x \rightarrow \infty} f(x) = l$.

(b) if

(i) $g(x) \leq f(x)$, $\forall x \in (R, \infty)$

(ii) $g(x) \rightarrow \infty$ as $x \rightarrow \infty$
 then $f(x) \rightarrow \infty$ as $x \rightarrow \infty$.

Example. Let's examine the asymptotic behaviour as $x \rightarrow \infty$ of:

$$f(x) = \frac{\sin(1/x)}{x} \quad (9.2.5)$$

We have that for $x \neq 0$:

$$-1 \leq \sin(1/x) \leq 1 \quad (9.2.6)$$

so that:

$$-\frac{1}{x} \leq \frac{\sin(1/x)}{x} \leq \frac{1}{x}, \quad \forall x \neq 0 \quad (9.2.7)$$

Then, since:

$$\lim_{x \rightarrow \infty} -\frac{1}{x} = \lim_{x \rightarrow \infty} \frac{1}{x} = 0 \quad (9.2.8)$$

we must have that

$$\lim_{x \rightarrow \infty} \frac{\sin(1/x)}{x} = 0 \quad (9.2.9)$$

◀

Proposition 9.11 (Asymptotic behaviour of standard functions)

(a) let $a_0, a_1, \dots, a_{n-1} \in \mathbb{R}$, and let

$$p(x) = x^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0 \quad (9.2.10)$$

Then

$$p(x) \rightarrow \infty, \quad \text{and} \quad \frac{1}{p(x)} \rightarrow 0 \quad \text{as } x \rightarrow \infty \quad (9.2.11)$$

(b) For $n \in \mathbb{N}$, then:

$$\frac{e^x}{x^n} \rightarrow \infty, \quad \text{and} \quad \frac{x^n}{e^x} \rightarrow 0 \quad \text{as } x \rightarrow \infty \quad (9.2.12)$$

(c) we have $\log x \rightarrow \infty$ as $x \rightarrow \infty$, and for $a > 0$ we have:

$$\lim_{x \rightarrow \infty} \frac{\log x}{x^a} = 0 \quad (9.2.13)$$

Proof. (a) Firstly, we have that the zeros of the polynomial p must lie in some interval $(-M, M)$, so we have that:

$$p(x) > 0, \quad \forall x \in (M, \infty) \quad (9.2.14)$$

Now if $x \neq 0$ then:

$$p(x) = x^n \left(1 + \frac{a_{n-1}}{x} + \dots + \frac{a_0}{x^n} \right) \quad (9.2.15)$$

Now we have that:

$$1 + \frac{a_{n-1}}{x} + \dots + \frac{a_0}{x^n} \rightarrow 1 + 0 + \dots + 0 = 1 \text{ as } x \rightarrow \infty \quad (9.2.16)$$

Therefore

$$\frac{1}{p(x)} = \frac{\frac{1}{x^n}}{1 + \frac{a_{n-1}}{x} + \dots + \frac{a_0}{x^n}} \rightarrow 0 \text{ as } x \rightarrow \infty \quad (9.2.17)$$

It follows from the reciprocal rule that, since $p(x) \rightarrow \infty$ as $x \rightarrow \infty$.

(b) We use the series expansion:

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!} \geq \frac{x^{n+1}}{(n+1)!}, \quad \forall x \geq 0 \quad (9.2.18)$$

Hence, if $x > 0$ then:

$$\frac{e^x}{x^n} \geq \frac{x}{(n+1)!}, \quad \text{and} \quad 0 \leq \frac{x^n}{e^x} \leq \frac{(n+1)!}{x} \quad (9.2.19)$$

Now $\frac{x}{(n+1)!} \rightarrow \infty$ so that:

$$\frac{e^x}{x^n} \rightarrow \infty, \quad \text{as } x \rightarrow \infty \quad (9.2.20)$$

Similarly, using the Squeeze rule

$$\frac{x^n}{e^x} \rightarrow 0 \text{ as } x \rightarrow \infty \quad (9.2.21)$$

(c) We know that $\log x$ is the strictly increasing inverse of the exponential, so that $\log x \rightarrow \infty$.

Now let $a > 0$, if we define $t(x) = a \log x$ then $x^a = e^{a \log x} = e^t$, giving:

$$\frac{\log x}{x^a} = \frac{t}{ae^t} \quad (9.2.22)$$

We have shown that $t \rightarrow \infty$ since a is positive. Hence, using part (b)

$$\frac{t}{ae^t} \rightarrow 0 \quad (9.2.23)$$

Using the composition rule of limits:

$$\lim_{x \rightarrow \infty} \frac{\log x}{x^a} = 0 \quad (9.2.24)$$

as desired.

Example. Let us examine the behaviour of

$$f(x) = \frac{2e^x - x^2}{e^x + \log x} \quad (9.2.25)$$

Then:

$$f(x) = \frac{\frac{2e^x}{x^2} - 1}{\frac{e^x}{x^2} + \frac{\log x}{x^2}} \quad (9.2.26)$$

Using the combination rules of limits:

$$\lim_{x \rightarrow \infty} f(x) = \frac{\lim_{x \rightarrow \infty} \frac{2e^x}{x^2} - 1}{\lim_{x \rightarrow \infty} \left(\frac{e^x}{x^2} + \frac{\log x}{x^2} \right)} \quad (9.2.27)$$

$$= \frac{\lim_{x \rightarrow \infty} \frac{2e^x}{x^2} - 1}{\lim_{x \rightarrow \infty} \frac{e^x}{x^2}} \quad (9.2.28)$$

$$= \lim_{x \rightarrow \infty} \left(2 - \frac{x^2}{e^x} \right) \quad (9.2.29)$$

$$= 2 \quad (9.2.30)$$

Example. Let's prove that $\lim_{x \rightarrow \infty} x \sin \frac{1}{x} = 1$. We use the substitution $u(x) = \frac{1}{x}$ so that:

$$\lim_{x \rightarrow \infty} x \sin \frac{1}{x} = \lim_{u \rightarrow 0} \frac{\sin u}{u} = 1 \quad (9.2.31)$$

as desired. ◀

9.3 Continuity of functions

Definition 9.12 (*Continuity of functions*)

Let f have a domain A and let $c \in A$. Then f is **continuous** at c if $\forall \epsilon > 0$, there exists $\delta > 0$ such that:

$$|f(x) - f(c)| < \epsilon, \quad \forall x \in A \text{ with } |x - c| < \delta \quad (9.3.1)$$

Much like in the definition of limits for series, we can view definition 9.12 as an $\epsilon - \delta$ game. One player chooses a small positive ϵ , and challenges the other player to find δ suitable small such that

$$|f(x) - f(c)| < \epsilon, \quad \forall x \in A \text{ with } |x - c| < \delta \quad (9.3.2)$$

is satisfied.

The general strategy to prove the continuity of polynomial functions f with domain A at some point $c \in A$ is:

- (i) express $f(x) - f(c) = (x - c)g(x)$
- (ii) obtain a bound $|g(x)| \leq M$ for $|x - c| \leq r$ where $r > 0$ such that $[c - r, c + r] \subset A$.
- (iii) use $|f(x) - f(c)| \leq M|x - c|$ for $|x - c| \leq r$ and set $\epsilon = M|x - c|$ to choose δ such that:

$$|f(x) - f(c)| < \epsilon, \forall x \in A \text{ with } |x - c| < \delta \quad (9.3.3)$$

Example. Let us prove that $f(x) = x^3$ is continuous at $c = 1$.

Firstly we note that the domain of f is \mathbb{R} . Suppose we are given $\epsilon > 0$, our goal is to choose $\delta > 0$ such that:

$$|x^3 - 1| < \epsilon, \forall x \text{ with } |x - 1| < \delta \quad (9.3.4)$$

We can write that:

$$x^3 - 1 = (x - 1)(x^2 + x + 1) \quad (9.3.5)$$

Now we find an upper bound for $|x^2 + x + 1|$. When $|x - 1| \leq 1$ then $x \in [0, 2]$ so that:

$$|x^2 + x + 1| \leq |x^2| + |x + 1| \leq 4 + 3 = 7, \forall |x - 1| \leq 1 \quad (9.3.6)$$

Therefore:

$$|f(x) - f(1)| \leq 7|x - 1|, \forall |x - 1| \leq 1 \quad (9.3.7)$$

Therefore, if $|x - 1| < \delta$, then $|f(x) - f(1)| \leq 7\delta|x - 1|$. Now we need $5\delta < \epsilon$ so that $\delta \leq \frac{1}{7}\epsilon$. Hence, we must have that if $\delta = \min\{1, \frac{\epsilon}{7}\}$ then:

$$|f(x) - f(1)| < \epsilon, \forall x \text{ with } |x - 1| < \delta \quad (9.3.8)$$

as desired. ◀

Theorem 9.13 (Equivalence of continuity definitions)

The $\epsilon - \delta$ definition and the sequence definition of continuity are equivalent.

Proof. Let f have domain A with $c \in A$.

Assume that continuity according to $\epsilon - \delta$ is satisfied, so that for $\epsilon > 0$, $\exists \delta > 0$ such that:

$$|f(x) - f(c)| < \epsilon, \text{ for } |x - c| < \delta \quad (9.3.9)$$

Now consider a sequence $x_n \in A$ such that $x_n \rightarrow c$. Then, there exists N such that:

$$|x_n - c| < \delta, \forall n > N \quad (9.3.10)$$

so that:

$$|f(x_n) - f(c)| < \epsilon, \forall n > N \quad (9.3.11)$$

Consequently, $f(x_n) \rightarrow f(c)$, which the sequence definition of continuity.

Now suppose that f is continuous according to the sequence definition, we argue by contradiction that for some $\epsilon > 0$, there is no $\delta > 0$ such that

$$|f(x) - f(c)| < \epsilon, \text{ for } |x - c| < \delta \quad (9.3.12)$$

Hence, for all n , $\exists x_n \in A$ with $|x_n - c| < \frac{1}{n}$ such that

$$|f(x) - f(c)| \geq \epsilon \quad (9.3.13)$$

By the sequential definition $\lim_{n \rightarrow \infty} f(x_n) = f(c)$ contradicting the above inequality. Hence the $\epsilon - \delta$ definition must also be satisfied. ■

9.4 Unusual function continuity

Proposition 9.14 (Dirichlet function)

The Dirichlet function defined as:

$$f(x) = \begin{cases} 1, & \text{if } x \text{ is rational} \\ 0, & \text{if } x \text{ is irrational} \end{cases} \quad (9.4.1)$$

is discontinuous everywhere on \mathbb{R} .

Proof. Let $c \in \mathbb{R}$. By the density of \mathbb{R} , each interval $(c - \frac{1}{n}, c + \frac{1}{n})$ with n natural contains a rational x_n and irrational y_n . Then $x_n \rightarrow c$ and $y_n \rightarrow c$, yet $f(x_n) = 1$ and $f(y_n) = 0$ so f is discontinuous at c . ■

Proposition 9.15 (Blancmange function)

The sawtooth function defined as:

$$s(x) = \begin{cases} x - \lfloor x \rfloor, & \text{if } 0 \leq x - \lfloor x \rfloor \leq \frac{1}{2} \\ 1 - (x - \lfloor x \rfloor), & \text{if } \frac{1}{2} < x - \lfloor x \rfloor < 1 \end{cases} \quad (9.4.2)$$

and the Blancmange function B is defined as:

$$B(x) = \sum_{n=0}^{\infty} \frac{1}{2^n} s(2^n x) \quad (9.4.3)$$

is continuous everywhere on \mathbb{R} .

Proof. Let $c \in \mathbb{R}$, and let $\epsilon > 0$. Then:

$$B(x) - B(c) = \sum_{n=0}^{\infty} \frac{1}{2^n} (s(2^n x) - s(2^n c)) \quad (9.4.4)$$

Using the triangle inequality:

$$|B(x) - B(c)| = \left| \sum_{n=0}^{\infty} \frac{1}{2^n} (s(2^n x) - s(2^n c)) \right| \leq \sum_{n=0}^{\infty} \frac{1}{2^n} |s(2^n x) - s(2^n c)| \quad (9.4.5)$$

Now since $s(2^n x) \in \left[0, \frac{1}{2}\right]$ and $s(2^n c) \in \left[0, \frac{1}{2}\right]$ for all x, c and natural n , we may write:

$$|s(2^n x) - s(2^n c)| \leq \frac{1}{2} \implies \sum_{n=N}^{\infty} \frac{1}{2^n} |s(2^n x) - s(2^n c)| \leq \frac{1}{2} \sum_{n=N}^{\infty} \frac{1}{2^n} \quad (9.4.6)$$

and using our standard results for geometric series:

$$\sum_{n=N}^{\infty} \frac{1}{2^n} |s(2^n x) - s(2^n c)| \leq \frac{1}{2^N} \quad (9.4.7)$$

Now consider:

$$x \mapsto s(2^n x), \quad n = 0, 1, 2, \dots \quad (9.4.8)$$

which is a continuous function. Consequently, for all n there is a positive δ_n such that:

$$|s(2^n x) - s(2^n c)| < \frac{\epsilon}{4}, \quad \forall |x - c| < \delta_n \quad (9.4.9)$$

Choosing $\delta = \min_{n \in \mathbb{N}} \delta_n$ we get that for $|x - c| < \delta$

$$\sum_{n=0}^{N-1} \frac{1}{2^n} |s(2^n x) - s(2^n c)| \leq \sum_{n=0}^{N-1} \frac{1}{2^N} \frac{\epsilon}{4} < 2 \cdot \frac{\epsilon}{4} = \frac{\epsilon}{2} \quad (9.4.10)$$

Consequently

$$\sum_{n=0}^{\infty} \frac{1}{2^n} |s(2^n x) - s(2^n c)| \leq \frac{\epsilon}{2} + \frac{1}{2^N} \quad (9.4.11)$$

so we need to choose N such that $\frac{1}{2^N} < \frac{1}{2}\epsilon$ for the condition of continuity to be satisfied. We can always do so because $\frac{1}{2^n}$ is a basic null sequence.

The blancmange function is therefore continuous ■

Definition 9.16 ($\epsilon - \delta$ definition of limit)

Let f be a function defined on $N_\epsilon(c)$ of c . Then $f(x)$ tends to l as x tends to c if $\forall \epsilon > 0, \exists \delta > 0$ such that

$$|f(x) - l| < \epsilon, \quad \forall x \text{ such that } 0 < |x - c| < \delta \quad (9.4.12)$$

We then write that:

$$\lim_{x \rightarrow c} f(x) = l \quad (9.4.13)$$

Example. We evaluate

$$\lim_{x \rightarrow 1} \frac{2x^3 + 3x - 5}{x - 1} \quad (9.4.14)$$

Since $2x^3 + 3x - 5 = (x - 1)(2x^2 + 2x + 5)$ we guess that the limit tends to 9. Indeed, we need to show that for each $\epsilon > 0$, there exists $\delta > 0$ such that:

$$|f(x) - 9| < \epsilon, \quad \forall x \text{ with } 0 < |x - 1| < \delta \quad (9.4.15)$$

But we have that for $0 < |x - 1| < 1$ then $x \in (-1, 2)$ so that:

$$|2x^2 + 2x + 5| < 17, \quad \forall 0 < |x - 1| < 1 \quad (9.4.16)$$

so that

$$|f(x) - 9| < 17|x - 1|, \quad \forall 0 < |x - 1| < 1 \quad (9.4.17)$$

Consequently if $|x - 1| < \delta$ then:

$$|f(x) - 9| < 17\delta, \quad \forall 0 < |x - 1| < \delta \quad (9.4.18)$$

To have that $|f(x) - 9| < \epsilon$, we need $\delta \leq \frac{\epsilon}{17}$. So given $\epsilon > 0$ we need to find $\delta \leq \frac{\epsilon}{17}$, which can always be done. Consequently f is continuous. \blacktriangleleft

9.5 Uniform continuity

Definition 9.17 (Uniform continuity)

A function f defined on the interval I is **uniformly continuous** on I if for all $\epsilon > 0$ then $\exists \delta > 0$ such that:

$$|f(x) - f(y)| < \epsilon, \quad \forall x, y \in I \text{ such that } |x - y| < \delta \quad (9.5.1)$$

Example. Let us prove that $f(x) = x^2$ is uniformly continuous on $I = [-4, 4]$.

Let $\epsilon > 0$, we have:

$$f(x) - f(y) = x^2 - y^2 = (x + y)(x - y) \quad (9.5.2)$$

Therefore $x, y \in [-4, 4]$ implies that $|x| \leq 4$ and $|y| \leq 4$ so that:

$$|f(x) - f(y)| = |x + y||x - y| \quad (9.5.3)$$

$$\leq (|x| + |y|)|x - y| \quad (9.5.4)$$

$$\leq 8|x - y| \quad (9.5.5)$$

Thus if we choose $\delta = \frac{\epsilon}{8}$ and $|x - y| < \delta$ then:

$$|f(x) - f(y)| < \epsilon \quad (9.5.6)$$

as desired. \blacktriangleleft

Theorem 9.17 (Sequential definition of uniform discontinuity)

Let f be defined on I , then f is not uniformly continuous on I iff $\exists (x_n), (y_n) \in I$ and $\exists \epsilon > 0$ such that:

- (i) $|x_n - y_n| \rightarrow 0$ as $n \rightarrow \infty$
- (ii) $|f(x_n) - f(y_n)| \geq \epsilon$, for $n = 1, 2, \dots$

Proof. Suppose f is not uniformly continuous on I . Then $\exists \epsilon > 0$ such that for all $\delta > 0$ there are $x, y \in I$ such that:

$$|x - y| < \delta \text{ and } |f(x) - f(y)| \geq \epsilon \quad (9.5.7)$$

Setting $\delta = 1, \frac{1}{2}, \frac{1}{3}, \dots$ we obtain

$$|x_n - y_n| < \frac{1}{n} \text{ and } |f(x_n) - f(y_n)| \geq \epsilon, \quad n = 1, 2, \dots \quad (9.5.8)$$

Therefore $|x_n - y_n| \rightarrow 0$ and $|f(x_n) - f(y_n)| \geq \epsilon$ as desired.

Now suppose that $|x_n - y_n| \rightarrow 0$ and $|f(x_n) - f(y_n)| \geq \epsilon$ are satisfied. Furthermore, suppose that f is uniformly continuous such that there exists $\delta > 0$ satisfying:

$$|f(x) - f(y)| < \epsilon, \quad \forall x, y \in I \text{ with } |x - y| < \delta \quad (9.5.9)$$

But by statement 1, $|x_n - y_n| < \delta$ for $n > N$, so

$$|f(x_n) - f(y_n)| < \epsilon, \quad \forall n > N \quad (9.5.10)$$

contradicting the second statement. Therefore f is not uniformly continuous on I . \blacksquare

Example. We show that $f(x) = x^2$ is not uniformly continuous on \mathbb{R} .
Indeed, taking $x_n = n + \frac{1}{n}$ and $y_n = n$ then we see that:

$$|x_n - y_n| = \left| \frac{1}{n} \right| = \frac{1}{n} \rightarrow 0 \quad (9.5.11)$$

Moreover:

$$|f(x_n) - f(y_n)| = \left| \left(n + \frac{1}{n} \right)^2 - n^2 \right| \quad (9.5.12)$$

$$= 2 + \frac{1}{n^2} > \epsilon \quad (9.5.13)$$

where $\epsilon = 2$.



Unit F2: Differentiation

Definition 10.1 (Differentiability) Let f be defined on the open interval I , and let $c \in I$. Then the **derivative of f at c** is defined as

$$f'(c) = \lim_{x \rightarrow c} \frac{f(x) - f(c)}{x - c} = \lim_{x \rightarrow c} Q(x) \quad (10.0.1)$$

where $Q(x)$ is the difference quotient. If this limit exists, then f is differentiable at c . If f is differentiable $\forall c \in I$ then we say that f is **differentiable**.

Example. Let us prove that $f(x) = \frac{1}{x}$ is differentiable on $I = \mathbb{R}^*$. Indeed the difference quotient reads:

$$Q(x) = \frac{f(x) - f(c)}{x - c} = \frac{1/x - 1/c}{x - c} = -\frac{1}{cx} \quad (10.0.2)$$

so that:

$$\lim_{x \rightarrow c} Q(x) = \lim_{x \rightarrow c} \frac{1}{cx} = \frac{1}{c^2} \quad (10.0.3)$$

where $c \in \mathbb{R}^*$. We may therefore conclude that:

$$f'(x) = -\frac{1}{x^2} \quad (10.0.4)$$



Example. Let us prove that the function:

$$f(x) = \begin{cases} x^2 \cos \frac{1}{x}, & x \neq 0 \\ 0, & x = 0 \end{cases} \quad (10.0.5)$$

is differentiable at $x = 0$. Indeed:

$$Q(x) = \frac{f(x) - f(0)}{x} = x \cos \frac{1}{x} \quad (10.0.6)$$

so we must prove that the following limit exists:

$$\lim_{x \rightarrow 0} x \cos \frac{1}{x} \quad (10.0.7)$$

Let us define

$$f_{\text{new}}(x) = \begin{cases} x \cos \frac{1}{x}, & x \neq 0 \\ 0, & x = 0 \end{cases} \quad (10.0.8)$$

To do so, we know that:

$$-x \leq x \cos \frac{1}{x} \leq x \quad (10.0.9)$$

Now by the squeeze rule, we see that since $h(x) = x$ and $g(x) = -x$ are continuous at 0 and $h(0) = g(0) = f_{\text{new}}(0) = 0$ then

$$\lim_{x \rightarrow 0} x \cos \frac{1}{x} = 0 \quad (10.0.10)$$

so that f is indeed differentiable at $x = 0$ with $f'(0) = 0$. ◀

Example. Let us examine the differentiability of

$$f(x) = \begin{cases} \sqrt{|x|} \sin \frac{1}{x}, & x \neq 0 \\ 0, & x = 0 \end{cases} \quad (10.0.11)$$

The difference quotient is:

$$Q(x) = \frac{\sqrt{|x|} \sin \frac{1}{x}}{x} \quad (10.0.12)$$

Now consider the sequence $h_n = \frac{1}{n\pi}$ so that

$$Q(h_n) = n\pi \sqrt{\frac{1}{n\pi}} \sin n\pi \quad (10.0.13)$$

which does not exist. Hence we must have that $\lim_{x \rightarrow 0} Q(x)$ does not exist. ◀

Proposition 10.2 (Standard derivatives)

- (i) $f(x) = k$ with $k \in \mathbb{R}$ then $f'(x) = 0$
- (ii) $f(x) = x^n$ with $n \in \mathbb{N}$ then $f'(x) = nx^{n-1}$
- (iii) $f(x) = \sin x$ then $f'(x) = \cos x$
- (iv) $f(x) = \cos x$ then $f'(x) = -\sin x$
- (v) $f(x) = e^x$ then $f'(x) = e^x$

Alternatively, in some situations it may be easier to define the difference quotient as:

$$Q(h) = \frac{f(h+c) - f(h)}{h} \quad (10.0.14)$$

so that differentiability at c is given when the following limit exists:

$$\lim_{h \rightarrow 0} Q(h) \quad (10.0.15)$$

Proof. (a) we have that

$$Q(x) = \frac{k-k}{x-c} = 0 \implies f'(x) = 0 \quad (10.0.16)$$

(b) we have that:

$$Q(h) = \frac{(c+h)^n - c^n}{h} = nc^{n-1} + \frac{n(n-1)}{2}c^{n-2}h + \dots + h^{n-1} \quad (10.0.17)$$

so that $Q(h) \rightarrow nc^{n-1}$ as $h \rightarrow 0$.

(c) The difference quotient is:

$$Q(h) = \frac{\sin(c+h) - \sin c}{h} = \frac{\sin c \cos h + \sin h \cos c - \sin c}{h} \quad (10.0.18)$$

$$= \cos c \frac{\sin h}{h} + \sin c \left(\frac{\cos h - 1}{h} \right) \quad (10.0.19)$$

so using some standard trigonometric limits:

$$\lim_{h \rightarrow 0} Q(h) = \cos c \implies f'(x) = \cos x \quad (10.0.20)$$

(d) The difference quotient is:

$$Q(h) = \frac{\cos(c+h) - \cos c}{h} = \frac{\cos c \cos h - \sin h \sin c - \cos c}{h} \quad (10.0.21)$$

$$= -\sin c \frac{\sin h}{h} + \cos c \left(\frac{\cos h - 1}{h} \right) \quad (10.0.22)$$

so using some standard trigonometric limits:

$$\lim_{h \rightarrow 0} Q(h) = -\sin c \implies f'(x) = -\sin x \quad (10.0.23)$$

(e) The difference quotient is:

$$Q(h) = \frac{e^c e^h - e^c}{h} = e^c \frac{e^h - 1}{h} \quad (10.0.24)$$

so $Q(h) \rightarrow e^c$ as $h \rightarrow 0$ proving that $f'(x) = e^x$.

Definition 10.3 (One-sided derivative)

Let f be defined on I and let $c \in I$. Then the **left derivative of f at c** is:

$$f'_L(c) = \lim_{x \rightarrow c^-} \frac{f(x) - f(c)}{x - c} = \lim_{h \rightarrow 0^-} Q(h) \quad (10.0.25)$$

If this limit exists, then f is left-differentiable at c . Similarly for the right derivative:

$$f'_R(c) = \lim_{x \rightarrow c^+} \frac{f(x) - f(c)}{x - c} = \lim_{h \rightarrow 0^+} Q(h) \quad (10.0.26)$$

Proposition 10.4 (Differentiability and one-sided differentiability)

Let f be defined on I and let $c \in I$.

$$f \text{ is differentiable at } c \iff f'_R(c) = f'_L(c) = f'(c). \quad (10.0.27)$$

Theorem 10.5 (Glue rule for differentiation)

Let f be defined on I and let $c \in I$. If there are functions g, h defined on I such that

- (1) $f(x) = g(x)$ for $x \in I, x < c$ and $f(x) = h(x)$ for $x \in I, x > c$
- (2) $f(c) = g(c) = h(c)$
- (3) g, h are differentiable at c

then f is differentiable at c iff $g'(c) = h'(c)$. In this case then $f'(c) = g'(c) = h'(c)$.

Note also that since differentiability is a local property, if we define a piece-wise function such as:

$$f(x) = \begin{cases} g(x), & x > c \\ h(x), & x \leq c \end{cases} \quad (10.0.28)$$

then we will have that:

$$f'(x) = \begin{cases} g'(x), & x > c \\ h'(x), & x < c \end{cases} \quad (10.0.29)$$

Example. Let us prove that:

$$f(x) = \begin{cases} -x^2, & x < 0 \\ x^2, & x \geq 0 \end{cases} \quad (10.0.30)$$

is differentiable on \mathbb{R} . Indeed define $g(x) = -x^2$ and $h(x) = x^2$ so that $f(x) = g(x)$

for $x \in \mathbb{R}, x < 0$ and $f(x) = h(x)$ for $x \in \mathbb{R}, x > 0$. Moreover, we also have that:

$$f(0) = g(0) = h(0) = 0 \quad (10.0.31)$$

Finally, we also know that $g'(0) = 0$ and $h'(0) = 0$, so that by the Glue rule $f'(0) = 0$.

$$f'(x) = \begin{cases} g'(x) = -2x, & x < 0 \\ h'(x) = 2x, & x > 0 \\ 0, & x = 0 \end{cases} = 2|x|, \quad x \in \mathbb{R} \quad (10.0.32)$$

as desired. ◀

10.1 Continuity and differentiability

Theorem 10.6 (*Differentiability implies continuity*)

Let f be defined on I , and let c . If f is differentiable at c then it is continuous at c .

Proof. Suppose f is differentiable at c so that

$$f'(c) = \lim_{x \rightarrow c} \frac{f(x) - f(c)}{x - c} \quad (10.1.1)$$

For $x \in I$ and $x \neq c$ we have that:

$$f(x) - f(c) = f'(c)(x - c) \quad (10.1.2)$$

Hence

$$\lim_{x \rightarrow c} f(x) - f(c) = f'(c) \cdot 0 = 0 \quad (10.1.3)$$

so that $f(x) \rightarrow f(c)$ implying continuity. ■

10.2 Rules of differentiation

Proposition 10.7 (*Combination rules*)

Let f, g be defined on I and let $c \in I$. If f, g are differentiable at c then:

- (i) $(f + g)'(c) = f'(c) + g'(c)$
- (ii) $(\lambda f)'(c) = \lambda f'(c)$ with $\lambda \in \mathbb{R}$
- (iii) $(fg)'(c) = f'(c)g(c) + f(c)g'(c)$
- (iv) if $g(c) \neq 0$ then $\left(\frac{f}{g}\right)'(c) = \frac{g(c)f'(c) - f(c)g'(c)}{(g(c))^2}$

Proof. (i) Let $F = f + g$, then:

$$\frac{F(x) - F(c)}{x - c} = \frac{f(x) - f(c)}{x - c} + \frac{g(x) - g(c)}{x - c} \quad (10.2.1)$$

$$\rightarrow f'(c) + g'(c) \quad (10.2.2)$$

as required.

(ii) Use product rule with $f = \lambda$.

(iii) Let $F = fg$ then:

$$\frac{F(x) - F(c)}{x - c} = \frac{f(x)g(x) - f(c)g(c)}{x - c} + \frac{g(x) - g(c)}{x - c} \quad (10.2.3)$$

$$= \frac{f(x) - f(c)}{x - c}g(c) + f(c)\frac{g(x) - g(c)}{x - c} \quad (10.2.4)$$

$$\rightarrow f'(c)g(c) + f(c)g'(c) \quad (10.2.5)$$

as required.

(iv) Let $F = \frac{f}{g}$, since g is continuous at c and $g(c) \neq 0$, there exists $\delta > 0$ such that $J = (c - \delta, c + \delta) \subseteq I$ and:

$$|g(x) - g(c)| < \frac{1}{2}|g(c)|, \quad \forall x \text{ with } |x - c| < \delta \quad (10.2.6)$$

This shows that there is some J such that $g(x) \neq 0$ for $x \in J$. Therefore, for $x \in J$ we have that:

$$\frac{F(x) - F(c)}{x - c} = \frac{f(x)/g(x) - f(c)/g(c)}{x - c} + \frac{g(x) - g(c)}{x - c} \quad (10.2.7)$$

$$= \frac{f(x)g(c) - f(c)g(x)}{(x - c)g(x)g(c)} \quad (10.2.8)$$

$$= \frac{g(c)(f(x) - f(c)) - f(c)(g(x) - g(c))}{(x - c)g(x)g(c)} \quad (10.2.9)$$

$$= \frac{1}{g(x)g(c)} \left(g(c) \frac{f(x) - f(c)}{x - c} - f(c) \frac{g(x) - g(c)}{x - c} \right) \quad (10.2.10)$$

$$\rightarrow \frac{g(c)f'(c) - f(c)g'(c)}{(g(c))^2} \quad (10.2.11)$$

as desired. ■

Theorem 10.8 (Composition rule) Let f be defined on I and let g be defined on J so that $f(I) \subset J$ and let $c \in I$.

If f is differentiable at c and g is differentiable at $f(c)$ then

$$(g \circ f)'(c) = g'(f(c))f'(c) \quad (10.2.12)$$

Proof. Let $F = g \circ f$ then:

$$\frac{F(x) - F(c)}{x - c} = \frac{g(f(x)) - g(f(c))}{x - c} \quad (10.2.13)$$

Let $y = f(x)$ with $x \in I$ and let $d = f(c)$. Then the right hand side of the above equation is:

$$\left(\frac{g(y) - g(d)}{y - d} \right) \left(\frac{f(x) - f(c)}{x - c} \right), \quad y \neq d \quad (10.2.14)$$

We may introduce the function $h(y)$ to deal with the discontinuity at $y \neq d$:

$$h(y) = \begin{cases} \frac{g(y) - g(d)}{y - d}, & y \neq d \\ g'(f(c)), & f(x) = f(c) \end{cases} \quad (10.2.15)$$

Since g is differentiable at d , we have $h(y) \rightarrow g'(d)$ as $y \rightarrow d$. Since $h(d) = g'(d)$ we have that h is continuous at d . We deduce that:

$$(h \circ f)(x) = \begin{cases} \frac{g(f(x)) - g(f(c))}{f(x) - f(c)}, & f(x) \neq f(c) \\ g'(f(c)), & f(x) = f(c) \end{cases} \quad (10.2.16)$$

is continuous at c .

Therefore:

$$\frac{F(x) - F(c)}{x - c} = (h \circ f)(x) \left(\frac{f(x) - f(c)}{x - c} \right) \quad (10.2.17)$$

so that as $x \rightarrow c$ then:

$$\frac{F(x) - F(c)}{x - c} \rightarrow g'(f(c))f'(c) \quad (10.2.18)$$

as desired. ■

Example. Let's find the derivative of $f(x) = \cos\left(\frac{\cos 2x}{x^2}\right)$.

Let us define

$$g(x) = \cos x, x \in I \quad (10.2.19)$$

and

$$h(x) = \frac{\cos 2x}{x^2}, x \in I = (0, \infty) \quad (10.2.20)$$

Then $h(I) \subseteq (0, \infty) = I$ so we may apply the composition rule:

$$(g \circ f)'(x) = -\sin\left(\frac{\cos 2x}{x^2}\right) \cdot \frac{-2x^2 \sin 2x - 2x \cos 2x}{x^4} \quad (10.2.21)$$

$$= 2 \sin\left(\frac{\cos 2x}{x^2}\right) \frac{x \sin 2x + \cos 2x}{x^3} \quad (10.2.22)$$

◀

Proposition 10.9 (*Inverse function rule*)

Let f be a function with domain I on which it is continuous and strictly monotonic. If it is differentiable on I and $f'(x) \neq 0$, $\forall x \in I$, then f^{-1} is differentiable on J . For $c \in I$ and $d = f(c)$ then:

$$(f^{-1})'(d) = \frac{1}{f'(c)} \quad (10.2.23)$$

Proof. We have that f is invertible on I with inverse f^{-1} whose domain is $J = f(I)$. Let $y \in J \setminus \{d\}$, it follows that $f^{-1}(y) = x \in I \setminus c$ due to the strict monotonicity of f . Therefore we find that:

$$\frac{f^{-1}(y) - f^{-1}(d)}{y - d} = \frac{x - c}{f(x) - f(c)} = \frac{1}{\frac{f(x) - f(c)}{x - c}} \quad (10.2.24)$$

Taking the limit as $y \rightarrow d \implies x = f^{-1}(y) \rightarrow c$ due to the continuity of f^{-1} . Hence:

$$\lim_{y \rightarrow d} \frac{f^{-1}(y) - f^{-1}(d)}{y - d} = \frac{1}{f'(c)} \quad (10.2.25)$$

proving that f^{-1} is differentiable at d with derivative $(f^{-1})'(d) = \frac{1}{f'(c)}$. ■

Example. Let us consider $f(x) = \tan x$, $x \in (-\pi/2, \pi/2)$. The domain of this function is $I = (-\pi/2, \pi/2)$, over which it is continuous and strictly increasing. Hence f will have an inverse f^{-1} with domain $f(I) = \mathbb{R}$.

Furthermore, f is differentiable on I , and its derivative is $f'(x) = \sec^2 x$, which is non-zero $\forall x \in I$. Therefore f^{-1} must be differentiable on \mathbb{R} by the Inverse function rule, with derivative at $y \in f(c)$, $\forall x \in (-\pi/2, \pi/2)$ given by:

$$(f^{-1})'(y) = \frac{1}{\sec^2 x} = \frac{1}{1 + \tan^2 x} = \frac{1}{1 + y^2} \quad (10.2.26)$$

implying that:

$$(\tan^{-1})'(x) = \frac{1}{1 + x^2}, \quad \forall x \in \mathbb{R} \quad (10.2.27)$$

◀

10.3 Rolle's theorem and local extrema

Definition (Local extrema)

The function f with domain J is said to have a:

1. **local maximum** $f(c)$ at $x = c$ if there exists $I = (c - r, c + r) \subseteq J$ where $r > 0$ such that:

$$f(x) \leq f(c), \quad \forall x \in I \quad (10.3.1)$$

2. **local minimum** $f(c)$ at $x = c$ if there exists $I = (c - r, c + r) \subseteq J$ where $r > 0$ such that:

$$f(c) \leq f(x), \quad \forall x \in I \quad (10.3.2)$$

3. **local extremum** $f(c)$ at $x = c$ if $f(c)$ is either a local maximum or minimum

Therefore, a local extremum is a value of f at some point such that f has lower values in some neighborhood of c .

Theorem (Local extreme value theorem)

If f has a local extremum at c and is differentiable at c then $f'(c) = 0$

Proof. Suppose that f has a local maximum at c so that $\exists r > 0$ such that:

$$f(x) \leq f(c), \quad \text{for } c - r < x < c + r \quad (10.3.3)$$

Let $x_n = c + \frac{r}{n}$ and $x'_n = c - \frac{r}{n}$ for $n = 2, 3, \dots$. Then $c < x_n < c + r$ so that $f(x_n) \leq f(c)$ and $x_n > c$. Hence:

$$\frac{f(x_n) - f(c)}{x_n - c} \leq 0 \implies f'(c) \leq 0 \quad (10.3.4)$$

We also have that $c - r < x'_n < c$ so that $f(x'_n) \leq f(c)$ and $x'_n < c$ and hence:

$$\frac{f(x'_n) - f(c)}{x_n - c} \geq 0 \implies f'(c) \geq 0 \quad (10.3.5)$$

We deduce that $f'(c) = 0$ as desired. ■

It is important to note that the converse of the local extreme value theorem is not necessarily true. All we know is that local extreme values of a differentiable function f on $[a, b]$ occur either at $x = a, x = b$ or at points $x \in (a, b)$ where $f'(x) = 0$.

Example. Let's find the local extrema of $f(x) = \sin^2 x + \cos x$ on $[0, \pi/2]$. Firstly f is continuous on $[0, \pi/2]$, so we have that $f(0) = 1$ and $f(\pi/2) = 1$. Moreover f is differentiable on $(0, \pi/2)$ with:

$$f'(x) = 2 \sin x \cos x - \sin x \quad (10.3.6)$$

which vanishes when:

$$2 \sin x \cos x = \sin x \implies \sin x = 0 \text{ or } \cos x = \frac{1}{2} \quad (10.3.7)$$

We see that $\sin x = 0 \implies x = \pi n, \forall n \in \mathbb{Z}$. Instead $\cos x = \frac{1}{2} \implies x = 2\pi n + \frac{\pi}{3}, \forall n \in \mathbb{Z}$. Since we're restricted to the interval $(0, \pi/2)$ we only consider $x = \frac{\pi}{3}$ where $f(x) = \frac{3}{4} + \frac{1}{2} = \frac{5}{4}$. We see that this provides the largest value of f compared to $x = 0, x = \frac{\pi}{2}$ and is therefore the local maximum. The local minimum, on the other hand, occurs at the endpoints $x = 0, x = \frac{\pi}{2}$ where $f(x) = 1$. \blacktriangleleft

Theorem (Rolle's theorem)

Let f be continuous on $[a, b]$ and differentiable on (a, b) . If $f(a) = f(b)$ then there exists c with $c \in (a, b)$ such that $f'(c) = 0$.

Proof. Suppose f is constant on $[a, b]$, then clearly $f'(x) = 0$ everywhere on (a, b) .

Suppose f is not constant on $[a, b]$. Since f is continuous it must have both a maximum and minimum on $[a, b]$. At least one of these must be different from $f(a) = f(b)$ since f is non-constant. Hence f has an extreme value for some $c \in (a, b)$. The extreme value theorem then shows that $f'(c) = 0$ as desired. \blacksquare

10.4 Mean value theorem

Theorem (Mean value theorem)

Let f be continuous on $[a, b]$ and differentiable on (a, b) . Then $\exists c \in (a, b)$ such that:

$$f'(c) = \frac{f(b) - f(a)}{b - a} \quad (10.4.1)$$

Proof. The gradient of the chord joining the points $(a, f(a))$ and $(b, f(b))$ is:

$$m = \frac{f(b) - f(a)}{b - a} \quad (10.4.2)$$

Consequently its equation will be $y = \frac{f(b) - f(a)}{b - a}(x - a) + f(a)$. Let us then define:

$$h(x) = f(x) - \frac{f(b) - f(a)}{b - a}(x - a) - f(a) \quad (10.4.3)$$

We then have that $h(a) = h(b) = 0$ and that h is continuous on $[a, b]$, differentiable on (a, b) . Consequently Rolle's theorem tells us that there exists some $c \in (a, b)$ for which:

$$h'(c) = f'(c) - \frac{f(b) - f(a)}{b - a} = 0 \implies f'(c) = \frac{f(b) - f(a)}{b - a} \quad (10.4.4)$$

as desired. ■

Example. Let's consider $f(x) = xe^x$ over the interval $(0, 2)$. Clearly f is differentiable on this interval, and continuous on $[0, 2]$ by the product rule. We also have that:

$$m = \frac{2e^2}{2} = e^2 \implies \exists c \in (0, 2) \text{ s.t. } f'(c) = e^2 \quad (10.4.5)$$

by the mean value theorem. ◀

Proposition (Increasing-Decreasing) Let f be continuous on I and differentiable on the interior J of I . Then if:

- (i) $f'(x) \leq 0, \forall x \in J$ then f is decreasing on I .
- (ii) $f'(x) \geq 0, \forall x \in J$ then f is increasing on I .

Proof. Let us take $x_1, x_2 \in I$ such that $x_1 < x_2$. Then since f satisfies the condition for the mean value theorem there exists $c \in (x_1, x_2)$ such that:

$$f'(c) = \frac{f(x_2) - f(x_1)}{x_2 - x_1} \quad (10.4.6)$$

If $f'(x) \leq 0 \forall x \in J$ then $f'(c) \leq 0$ and hence $f(x_2) - f(x_1) \leq 0$ proving that f is decreasing on I .

If $f'(x) \geq 0 \forall x \in J$ then $f'(c) \geq 0$ and hence $f(x_2) - f(x_1) \geq 0$ proving that f is increasing on I . ■

This gives us an efficient way to prove inequalities. Indeed, suppose we wished to prove that

$$g(x) \geq h(x), \forall x \in [a, b] \quad (10.4.7)$$

Then we let $f(x) = g(x) - h(x)$, and if it is continuous on $[a, b]$ and differentiable on (a, b) we show that either:

$$f(a) \geq 0 \text{ and } f'(x) \geq 0 \forall x \in (a, b) \quad (10.4.8)$$

which shows that f is smallest at a in $[a, b]$, or:

$$f(b) \geq 0 \text{ and } f'(x) \leq 0 \forall x \in (a, b) \quad (10.4.9)$$

which shows that f is smallest at b in $[a, b]$. In both cases we have that f will always be positive, and hence that $g(x) \geq h(x)$ on $[a, b]$.

Example. Let's prove that for $\alpha \geq 1$ and $x \geq -1$:

$$(1+x)^\alpha \geq 1 + \alpha x \quad (10.4.10)$$

The case where $\alpha = 1$ is clearly true, so we assume that $\alpha > 1$.

Let us define $f(x) = (1+x)^\alpha - 1 - \alpha x$ for $x \in I = [-1, \infty)$ which is continuous on I and differentiable on its interior. The derivative of f over I is:

$$f'(x) = \alpha(1+x)^{\alpha-1} - \alpha = \alpha((1+x)^{\alpha-1} - 1) \quad (10.4.11)$$

We see that for $-1 < x < 0$ then $0 < 1+x < 1$ so that $0 < (1+x)^{\alpha-1} < 1$ and hence $f'(x) < 0$, f is decreasing on $(-1, 0)$.

Similarly for $0 < x$ then $1 < 1+x$ so that $1 < (1+x)^{\alpha-1}$ and hence $f'(x) > 0$, f is increasing on $(0, \infty)$.

Finally, $f(0) = 0 \geq 0$, from which it follows that $f(x) \geq 0$ for all $x \in [-1, \infty)$, as desired. \blacktriangleleft

Theorem (Second derivative test)

Let f be a twice-differentiable function defined on the open interval I containing c , such that $f'(c) = 0$ and f'' is continuous at c .

- (i) if $f''(c) > 0$ then $f(c)$ is a local minimum of f .
- (ii) if $f''(c) < 0$ then $f(c)$ is a local maximum of f .

Proof. Suppose that $f''(c) > 0$. Since f'' is continuous at c , we have that $\exists \delta > 0$ such that $(c - \delta, c + \delta) \subseteq I$ and:

$$|f''(x) - f''(c)| < \epsilon = \frac{1}{2}f''(c), \quad \forall x \in (c - \delta, c + \delta) \quad (10.4.12)$$

implying that $f''(x) > \frac{1}{2}f''(c) > 0$. Hence f' is strictly increasing on $(c - \delta, c + \delta)$. Furthermore we have that $f'(c) = 0$ so:

$$f'(x) < 0, \quad \forall x \in (c - \delta, c) \quad f'(x) > 0, \quad \forall x \in (c, c + \delta) \quad (10.4.13)$$

This implies that

$$f(x) \text{ is decreasing, } \forall x \in (c - \delta, c) \quad f(x) \text{ is increasing, } \forall x \in (c, c + \delta) \quad (10.4.14)$$

proving that f has a local minimum at c . \blacksquare

10.5 L'Hopital's rule

Theorem (Cauchy's mean value theorem)

Let f, g be continuous on $[a, b]$ and differentiable on (a, b) . Then $\exists c \in (a, b)$ such that:

$$f'(c)(g(b) - g(a)) = g'(c)(f(b) - f(a)) \quad (10.5.1)$$

Proof. Let us define:

$$h(x) = f(x)(g(b) - g(a)) - g(x)(f(b) - f(a)) \quad (10.5.2)$$

Clearly by the combination rules of continuity and differentiability, h must be continuous on $[a, b]$ and differentiable on (a, b) . Note also that:

$$h(a) = f(a)g(b) - g(a)f(b) = h(b) \quad (10.5.3)$$

We may therefore apply Rolle's Theorem:

$$\exists c \in (a, b) \text{ s.t. } h'(c) = f'(c)(g(b) - g(a)) - g'(c)(f(b) - f(a)) = 0 \quad (10.5.4)$$

thus implying that:

$$f'(c)(g(b) - g(a)) = g'(c)(f(b) - f(a)) \quad (10.5.5)$$

as desired. ■

Theorem (L'Hopital's rule)

Let f, g be differentiable on I which contains c , and suppose $f(c) = g(c) = 0$. Then:

$$\lim_{x \rightarrow c} \frac{f(x)}{g(x)} = \lim_{x \rightarrow c} \frac{f'(x)}{g'(x)} \quad (10.5.6)$$

provided the latter limit exists.

Proof. Assume that:

$$\lim_{x \rightarrow c} \frac{f'(x)}{g'(x)} = l \quad (10.5.7)$$

Let $\epsilon > 0$, then it follows from the existence of the above limit that $\exists \delta > 0$ such that:

$$\left| \frac{f'(x)}{g'(x)} - l \right| < \epsilon, \quad 0 < |x - c| < \delta \quad (10.5.8)$$

so we cannot have that $g'(x) = 0$ within $0 < |x - c| < \delta$. Note however that if $g(x_0) = g(c)$ for some x_0 (wlog $x_0 > c$), then by Rolle's theorem $g'(x) = 0$ for some $x \in (c, x_0)$, which is impossible. Hence we must have that $g(x) \neq g(c)$ for all $0 < |x - c| < \delta$. Applying

Cauchy's mean value theorem, we have some $d \in (c, x)$ such that:

$$f'(d)(g(x) - g(c)) = g'(d)(f(x) - f(c)) \implies \frac{f'(d)}{g'(d)} = \frac{f(x) - f(c)}{g(x) - g(c)} \quad (10.5.9)$$

and since $f(c) = g(c) = 0$ we get that:

$$\frac{f'(d)}{g'(d)} = \frac{f(x)}{g(x)} \quad (10.5.10)$$

Therefore:

$$\left| \frac{f(x)}{g(x)} - l \right| = \left| \frac{f'(d)}{g'(d)} - l \right| < \epsilon, \quad 0 < |x - c| < \delta \quad (10.5.11)$$

proving that

$$\lim_{x \rightarrow c} \frac{f(x)}{g(x)} = l \quad (10.5.12)$$

as desired. ■

Example. Consider

$$\lim_{x \rightarrow 0} \frac{\sin x - x \cos x}{x^3} \quad (10.5.13)$$

We see that $f(x) = \sin x - x \cos x$ is continuous and differentiable in \mathbb{R} by the combination rules, and so is $g(x) = x^3$. Moreover, we also have that $f(0) = g(0) = 0$, hence we may apply l'Hopital's theorem:

$$\lim_{x \rightarrow 0} \frac{\sin x - x \cos x}{x^3} = \lim_{x \rightarrow 0} \frac{\cos x - \cos x + x \sin x}{3x^2} = \lim_{x \rightarrow 0} \frac{\sin x}{3x} = \frac{1}{3} \quad (10.5.14)$$



Unit F3: Integration

11.1 The Riemann integral

Definition (*Partition*)

A partition P of a closed interval $[a, b]$ is a collection of closed non-intersecting subintervals whose union gives $[a, b]$:

$$P = \{[x_0, x_1], [x_1, x_2], \dots, [x_{i-1}, x_i], \dots, [x_{n-1}, x_n]\} \quad (11.1.1)$$

with:

$$a = x_0 < x_1 < \dots < x_i < \dots < x_n = b \quad (11.1.2)$$

The points x_i are known as partition points, and the i th subinterval is $I_i = [x_{i-1}, x_i]$ whose length is $\delta x_i = x_i - x_{i-1}$. Instead the mesh of P is defined as $\|P\| = \max_{1 \leq i \leq n} \{\delta x_i\}$.

Example. Consider the following partition P of $[0, 1]$:

$$P = \left\{ \left[0, \frac{1}{2}\right], \left[\frac{1}{2}, \frac{3}{5}\right], \left[\frac{3}{5}, \frac{3}{4}\right], \left[\frac{3}{4}, 1\right] \right\} \quad (11.1.3)$$

We find that:

$$\delta x_1 = \frac{1}{2}, \quad \delta x_2 = \frac{1}{10}, \quad \delta x_3 = \frac{3}{20}, \quad \delta x_4 = \frac{1}{4} \quad (11.1.4)$$

so that:

$$\|P\| = \max\{\delta x_1, \delta x_2, \delta x_3, \delta x_4\} = \frac{1}{2} \quad (11.1.5)$$



Definition (*Riemann sums*)

Let f be bounded on $[a, b]$ and let $P = \{[a, x_1], [x_1, x_2], \dots, [x_{i-1}, x_i], \dots, [x_{n-1}, b]\}$.

Define:

$$m_i = \inf_{x \in I_i} f, M_i = \sup_{x \in I_i} f \quad (11.1.6)$$

Then the lower Riemann sum for f on $[a, b]$ is:

$$L(f, P) = \sum_{i=1}^n m_i \delta x_i \quad (11.1.7)$$

while the upper Riemann sum for f on $[a, b]$ is:

$$U(f, P) = \sum_{i=1}^n M_i \delta x_i \quad (11.1.8)$$

Geometrically, the upper Riemann sum represents an upper bound for the area under f , whereas the lower Riemann sum represents a lower bound for the area under f .

At the essence of Riemann integration is that we may approximate a function as constant over a sufficiently small interval $I_i = [x_{i-1}, x_i]$. Doing this for the entire partition of P , we get a series of intervals over which f is taken to be constant. In other words, we may consider the area under f as a series of vertical strips of width δx_i . But what constant value should we take for the height of the columns? Well, m_i gives the largest lower bound of f over I_i , while M_i gives the smallest upper bound of f . Consequently $m_i \delta x_i$ will underestimate the area of the i th vertical strip, while $M_i \delta x_i$ will overestimate it.

Example. Consider the function

$$f(x) = \begin{cases} 2x, & 0 < x < 1 \\ 1, & x = 0, 1 \end{cases} \quad (11.1.9)$$

and the partition $P = \left\{ [0, \frac{1}{4}], [\frac{1}{4}, \frac{1}{2}], [\frac{1}{2}, \frac{3}{4}], [\frac{3}{4}, 1] \right\}$. Then we see that:

$$m_1 = 0, M_1 = 1, \delta x_1 = \frac{1}{4} \quad (11.1.10)$$

$$m_2 = \frac{1}{2}, M_2 = 1, \delta x_2 = \frac{1}{4} \quad (11.1.11)$$

$$m_3 = 1, M_3 = \frac{3}{2}, \delta x_3 = \frac{1}{4} \quad (11.1.12)$$

$$m_4 = 1, M_4 = 2, \delta x_4 = \frac{1}{4} \quad (11.1.13)$$

$$(11.1.14)$$

and therefore:

$$L(f, P) = \frac{1}{4}(0 + \frac{1}{2} + 1 + 1) = \frac{5}{8} \quad (11.1.15)$$

$$U(f, P) = \frac{1}{4}(1 + 1 + \frac{3}{2} + 2) = \frac{11}{8} \quad (11.1.16)$$

◀

Example. Let

$$f(x) = \begin{cases} x^2, & 0 \leq x \leq 1 \\ 2, & 1 < x \leq 2 \end{cases} \quad (11.1.17)$$

and the partition $P = \{[0, \frac{1}{n}], [\frac{1}{n}, \frac{2}{n}], \dots, [2 - \frac{1}{n}, 2]\}$.

Now we see that the i th interval is $[\frac{i-1}{n}, \frac{i}{n}]$, while the interval width is $\delta x_i = \frac{1}{n}$. Also, note that f is increasing over $[0, 2]$, so m_i will be the value of f at the left endpoint of the i th interval, while M_i will be the value of f at the right endpoint of the i th interval. Consequently, for $1 \leq i \leq n$ we find that $m_i = \frac{(i-1)^2}{n^2}$, while $M_i = \frac{i^2}{n^2}$. Instead for $i = n+1$, $m_i = 1$ (which is coherent with $m_i = \frac{(i-1)^2}{n^2}$) whereas $M_i = 2$. Finally for $n+2 \leq i \leq 2n$ we get that $m_i = M_i = 2$. Hence:

$$L(f, P) = \sum_{i=1}^{n+1} \frac{(i-1)^2}{n^2} \frac{1}{n} + \sum_{i=n+2}^{2n} 2 \frac{1}{n} \quad (11.1.18)$$

$$= \frac{1}{n^3} \sum_{i=0}^n i^2 + \sum_{i=n+2}^{2n} 2 \frac{1}{n} \quad (11.1.19)$$

$$= \frac{1}{n^3} \frac{n(n+1)(2n+1)}{6} + \frac{2}{n} (2n - n - 1) \quad (11.1.20)$$

$$= \frac{n(n+1)(2n+1)}{6n^3} + 2 - \frac{2}{n} \quad (11.1.21)$$

Instead:

$$U(f, P) = \sum_{i=1}^n \frac{i^2}{n^2} \frac{1}{n} + \sum_{n+1}^{2n} 2 \frac{1}{n} \quad (11.1.22)$$

$$= \frac{n(n+1)(2n+1)}{6n^3} + 2 \frac{1}{n} (2n - n) = \frac{n(n+1)(2n+1)}{6n^3} + 2 \quad (11.1.23)$$

Note that taking the limit as $n \rightarrow \infty$:

$$\lim_{n \rightarrow \infty} L(f, P) = \frac{1}{3} + 2 = \frac{7}{3} \quad (11.1.24)$$

and similarly:

$$\lim_{n \rightarrow \infty} U(f, P) = \frac{1}{3} + 2 = \frac{7}{3} \quad (11.1.25)$$

We define this number as the integral of f on $[a, b]$. ◀

Theorem (Lower and upper Riemann sum inequality)

Let f be bounded on $[a, b]$ and let P, P' be partitions of $[a, b]$. Then $L(f, P) \leq U(f, P')$.

Proof. Suppose f is non-negative on $[a, b]$.

Since for any given partition P , $m_i \leq M_i$ we have that $L(f, P) \leq U(f, P)$. Also, let P'' be the union of the P and P' partitions. For example, if $P = \left\{ [0, \frac{1}{4}], [\frac{1}{4}, \frac{1}{2}] \right\}$ and $P' = \left\{ [0, \frac{1}{3}], [\frac{1}{3}, \frac{1}{2}] \right\}$, then $P'' = \left\{ [0, \frac{1}{4}], [\frac{1}{4}, \frac{1}{3}], [\frac{1}{3}, \frac{1}{2}] \right\}$.

Now consider what happens when we add a new partition x' to $P = \{[x_0, x_1], \dots, [x_{n-1}, x_n]\}$. Since f is positive adding this partition will not increase the upper Riemann sum. Similarly, adding the partition will not decrease the lower Riemann sum.

In other words, if we add a refinement from the P partition onto the P' partition to form the P'' partition, we will find that $U(f, P'') \leq U(f, P')$. Similarly if we add a refinement from the P' partition onto the P partition to form the P'' partition, we will find that $L(f, P) \leq L(f, P'')$.

Therefore, we may write that:

$$L(f, P) \leq L(f, P'') \leq U(f, P'') \leq U(f, P') \quad (11.1.26)$$

as desired.

If instead f is negative over some interval, then because it is bounded we may still form the function $g = f + c$ where c is some constant. Applying the same reasoning as before we will find that $L(g, P) \leq U(g, P')$ and thus $L(f, P) \leq U(f, P')$ since the upper and lower riemann sums of a constant function c are identical. ■

Definition (Integral)

Let f be a bounded function on $[a, b]$, and let P be a partition of $[a, b]$. Then the lower integral of f on $[a, b]$ is:

$$\underline{\int_a^b} f = \sup_P L(f, P) \quad (11.1.27)$$

while the upper integral of f is:

$$\overline{\int_a^b} f = \inf_P U(f, P) \quad (11.1.28)$$

If these two are equal, then we say that f is integrable on $[a, b]$. The values they are equal to is the integral of f on $[a, b]$.

In the previous example, we would write that:

$$\underline{\int_0^2} f = \overline{\int_0^2} f = \frac{7}{3} \implies \int_0^2 f = \frac{7}{3} \quad (11.1.29)$$

Example. Let f be the Dirichlet function defined by

$$f(x) = \begin{cases} 1, & 0 \leq x \leq 1, \text{ } x \text{ is irrational} \\ 0, & 0 \leq x \leq 1, \text{ } x \text{ is rational} \end{cases} \quad (11.1.30)$$

with a partition $P = \{[0, x_1], [x_1, x_2], \dots, [x_{n-1}, 1]\}$. Due to the density of real numbers, we can always find a rational and irrational number within each interval, so that $m_i = 0$ and $M_i = 1$. Hence we find that:

$$L(f, P) = \sum_i m_i \delta x_i = 0 \quad (11.1.31)$$

while

$$U(f, P) = \sum_i M_i \delta x_i = \sum_i \delta x_i = 1 \quad (11.1.32)$$

Therefore:

$$\underline{\int_0^1} f = 0 \neq \overline{\int_0^1} f = 1 \quad (11.1.33)$$

from which it follows that f is not integrable on $[0, 1]$. ◀

Theorem (*Integrability*)

Let f be bounded on $[a, b]$, if there exists a sequence of partitions (P_n) of $[a, b]$ such that $\|P_n\| \rightarrow 0$ then:

$$\lim_{n \rightarrow \infty} L(f, P_n) = \lim_{n \rightarrow \infty} U(f, P_n) = A \in \mathbb{R} \quad (11.1.34)$$

then $\int_a^b f = A$.

Proof. Let $\epsilon > 0$, then there exists n such that:

$$|L(f, P_n) - A| < \frac{1}{2}\epsilon \implies L(f, P_n) > A - \frac{1}{2}\epsilon \quad (11.1.35)$$

and similarly:

$$U(f, P_n) < A + \frac{1}{2}\varepsilon \quad (11.1.36)$$

Also, by definition we must have that:

$$L(f, P_n) \leq \underline{\int_a^b} f \leq \overline{\int_a^b} f \leq U(f, P_n) \quad (11.1.37)$$

so that:

$$A - \frac{1}{2}\varepsilon \leq \underline{\int_a^b} f \leq \overline{\int_a^b} f \leq A + \frac{1}{2}\varepsilon \quad (11.1.38)$$

Therefore, we find that f is integrable on $[a, b]$ with:

$$\int_a^b f = A \quad (11.1.39)$$

as desired. ■

Proposition (Integrability)

A function f which is

- (a) bounded and monotonic on $[a, b]$ or
- (b) continuous on $[a, b]$

is integrable on $[a, b]$.

Proof. (a) Consider the following partition of $[a, b]$ into equal-sized intervals:

$$P_n = \{[a, x_1], [x_1, x_2], \dots, [x_{n-1}, n]\} \quad (11.1.40)$$

with $x_i = a + \frac{b-a}{n}i$ and thus $\delta x_i = \frac{b-a}{n}$. Since f is increasing, we must have that $m_i = f(x_{i-1})$ and $M_i = f(x_i)$. Hence:

$$U(f, P_n) - L(f, P_n) = \sum_i (f(x_i) - f(x_{i-1})) \frac{b-a}{n} = (f(b) - f(a)) \frac{b-a}{n} \quad (11.1.41)$$

This sequence is clearly null, so f is integrable on $[a, b]$. ■

Proposition (Properties of Riemann integral)

Let f be integrable on an interval containing a, b, c , then:

$$\int_a^c f + \int_c^b f = \int_a^b f \quad (11.1.42)$$

Suppose that f is integrable on $[a, b]$, then $|f|$ is also integrable on $[a, b]$. Also:

$$\int_a^b (f + g) = \int_a^b f + \int_a^b g \quad (11.1.43)$$

and

$$\int_a^b \lambda f = \lambda \int_a^b f \quad (11.1.44)$$

Finally, both fg and f/g are integrable, provided that $\frac{1}{g}$ is bounded on $[a, b]$ in the latter case.

11.2 Inequalities and series with integrals

Series

Proposition (Inequality rules)

Let f and g be integrable over $[a, b]$. then:

- (i) if $f(x) \leq g(x)$, $\forall x \in [a, b]$ then:

$$\int_a^b f \leq \int_a^b g \quad (11.2.1)$$

- (ii) if $m \leq f(x) \leq M$, $\forall x \in [a, b]$ then:

$$m(b - a) \leq \int_a^b f \leq M(b - a) \quad (11.2.2)$$

- (iii)

$$\left| \int_a^b f \right| \leq \int_a^b |f| \quad (11.2.3)$$

Proof. (i) Suppose that $f(x) \leq g(x)$, $\forall x \in [a, b]$, and let P be any partition of $[a, b]$.

Then:

$$\inf_{[x_i, x_{i+1}]} f \leq \inf_{[x_i, x_{i+1}]} g \quad (11.2.4)$$

implying that:

$$\int_a^b f = \sup_P L(f, p) \leq \sup_P L(g, P) = \int_a^b g \quad (11.2.5)$$

as desired.

- (ii) Suppose $m \leq f(x) \leq M$ over $[a, b]$. Then from the results of part (a)

$$\int_a^b m \leq \int_a^b f \leq \int_a^b M \implies m(b - a) \leq \int_a^b f \leq M(b - a) \quad (11.2.6)$$

(iii) Note that $-f(x) \leq |f(x)| \leq f(x)$ for all $x \in [a, b]$, so that:

$$-\int_a^b |f| \leq \int_a^b f \leq \int_a^b |f| \implies \left| \int_a^b f \right| \leq \int_a^b |f| \quad (11.2.7)$$

as desired. ■

Example.

(a) Let us prove that:

$$\int_1^3 x \sin \frac{1}{x^{10}} dx \leq 4 \quad (11.2.8)$$

We have that for all $x \in [1, 3]$:

$$-1 \leq \sin \frac{1}{x^{10}} \leq 1 \quad (11.2.9)$$

implying that:

$$\int_1^3 x \sin \frac{1}{x^{10}} dx \leq \int_1^3 x = \frac{1}{2}[x^2]_0^3 = 4 \quad (11.2.10)$$

as desired.

(b) Let us prove that:

$$\frac{1}{2} \leq \int_0^{\frac{1}{2}} e^{x^2} dx \leq \frac{1}{2} e^{1/4} \quad (11.2.11)$$

Indeed, note that:

$$\frac{d}{dx}(e^{x^2}) = 2xe^{x^2} \geq 0, \quad \forall x \in \left[0, \frac{1}{2}\right] \quad (11.2.12)$$

showing that e^{x^2} is increasing on $[0, \frac{1}{2}]$. It then follows that:

$$1 \leq e^{x^2} \leq e^{\frac{1}{4}}, \quad \forall x \in \left[0, \frac{1}{2}\right] \quad (11.2.13)$$

and hence:

$$\frac{1}{2} \leq \int_0^{\frac{1}{2}} e^{x^2} dx \leq \frac{1}{2} e^{1/4} \quad (11.2.14)$$

as desired.

(c) Finally, let us prove that:

$$\left| \int_0^{\frac{\pi}{4}} \frac{\tan x}{3 - \sin x^2} dx \right| \leq \frac{1}{4} \log 2 \quad (11.2.15)$$

Indeed, we have that:

$$-1 \leq \sin(x^2) \leq 1 \implies 2 \leq 3 - \sin x^2 \leq 4 \quad (11.2.16)$$

Thus:

$$\frac{1}{4} \tan x \leq \frac{\tan x}{3 - \sin x^2} \leq \frac{1}{2} \tan x \quad (11.2.17)$$

implying that:

$$\left| \frac{\tan x}{3 - \sin x^2} \right| \leq \frac{1}{2} |\tan x| \quad (11.2.18)$$

Also $\tan x \geq 0$, $\forall x \in [0, \frac{\pi}{4}]$, so that:

$$\left| \frac{\tan x}{3 - \sin x^2} \right| \leq \frac{1}{2} \tan x \quad (11.2.19)$$

Using the limit inequality:

$$\left| \int_0^{\frac{\pi}{4}} \frac{\tan x}{3 - \sin x^2} dx \right| \leq \int_0^{\frac{\pi}{4}} \frac{1}{2} \tan x dx = \frac{1}{2} [\ln |\sec x|]_0^{\frac{\pi}{4}} \quad (11.2.20)$$

$$= \frac{1}{2} \ln \left| \sqrt{2} \right| = \frac{1}{4} \ln 2 \quad (11.2.21)$$

as desired.



Wallis' formula

Lemma. Let

$$I_n = \int_0^{\frac{\pi}{2}} \sin^n x dx, \quad n = 0, 1, 2, \dots \quad (11.2.22)$$

Then $I_n = \frac{n-1}{n} I_{n-2}$ for $n \geq 2$.

Proof. It is easy to see that:

$$I_0 = \frac{\pi}{2}, \quad I_1 = 1 \quad (11.2.23)$$

Also, for $n \geq 2$:

$$I_n = \int_0^{\pi/2} \sin^n x dx = \int_0^{\pi/2} \sin x \sin^{n-1} x dx \quad (11.2.24)$$

$$= [-\cos x \sin^{n-1} x]_0^{\pi/2} + \int_0^{\pi/2} (n-1) \cos^2 x \sin^{n-2} x dx \quad (11.2.25)$$

$$= (n-1) \left(\int_0^{\pi/2} \sin^{n-2} x dx - \int_0^{\pi/2} \sin^n x dx \right) \quad (11.2.26)$$

$$\implies nI_n = (n-1)I_{n-2} \implies I_n = \frac{n-1}{n} I_{n-2} \quad (11.2.27)$$

as desired. ■

For example, we have that even powers of $\sin x$ integrate to:

$$I_4 = \frac{3}{4} \frac{1}{2} \frac{\pi}{2}, \quad I_6 = \frac{5}{6} \frac{3}{4} \frac{1}{2} \frac{\pi}{2} \quad (11.2.28)$$

or more generally:

$$I_{2n} = \frac{2n-1}{2n} \frac{2n-3}{2n-2} \cdots \frac{3}{4} \frac{1}{2} \frac{\pi}{2} \quad (11.2.29)$$

Similarly, we have that odd powers of $\sin x$ integrate to:

$$I_5 = \frac{4}{5} \frac{2}{3}, \quad I_7 = \frac{6}{7} \frac{4}{5} \frac{2}{3} \quad (11.2.30)$$

or more generally:

$$I_{2n+1} = \frac{2n}{2n+1} \frac{2n-2}{2n-1} \cdots \frac{4}{5} \frac{2}{3} \quad (11.2.31)$$

Lemma. Let:

$$a_n = \frac{2}{1} \frac{2}{3} \frac{4}{5} \frac{4}{7} \cdots \frac{2n}{2n-1} \frac{2n}{2n+1} \quad (11.2.32)$$

$$b_n = \frac{(n!)^2 2^{2n}}{(2n)! \sqrt{n}} \quad (11.2.33)$$

then:

$$b_n^2 = \frac{2n+1}{n} a_n, \quad n = 1, 2, \dots \quad (11.2.34)$$

Proof. Firstly note that:

$$b_1^2 = \frac{2^4}{4} = 4, \quad a_1 = \frac{2}{1} \frac{2}{3} = \frac{4}{3} \implies b_1^2 = 3a_1 \quad (11.2.35)$$

More generally, suppose that for some n :

$$b_n^2 = \frac{2n+1}{n} a_n \quad (11.2.36)$$

then:

$$b_{n+1}^2 = \frac{((n+1)!)^4 2^{4(n+1)}}{((2n+2)!)^2 (n+1)} \quad (11.2.37)$$

$$= \frac{(n!)^4 (n+1)^4 2^{4n} 2^4}{(2n!)^2 (2n+2)^2 (2n+1)^2 (n+1)} \quad (11.2.38)$$

$$= \frac{(n!)^4 2^{4n}}{((2n)!)^2 n} \frac{n}{n+1} \frac{(n+1)^4 2^4}{2^2 (n+1)^2 (2n+1)^2} \quad (11.2.39)$$

$$= \frac{2n+1}{n} \frac{n}{n+1} \frac{(n+1)^2 2^2}{(2n+1)^2} a_n \quad (11.2.40)$$

$$= \frac{4(n+1)}{2n+1} a_n \quad (11.2.41)$$

Now note that:

$$a_{n+1} = \frac{2n+2}{2n+1} \frac{2n+2}{2n+3} a_n \implies \frac{4(n+1)}{2n+1} a_n = \frac{2n+3}{n+1} a_{n+1} \quad (11.2.42)$$

and hence:

$$b_{n+1}^2 = \frac{2n+3}{n+1} a_{n+1} \quad (11.2.43)$$

as desired. ■

Theorem (Wallis' Formula) Wallis' formula:

$$\lim_{n \rightarrow \infty} \frac{(n!)^2 2^{2n}}{(2n)! \sqrt{n}} = \sqrt{\pi} \quad (11.2.44)$$

Proof. We begin by proving that:

$$\lim_{n \rightarrow \infty} a_n = \frac{\pi}{2} \quad (11.2.45)$$

Indeed, note that:

$$\frac{\pi}{2} a_n = \frac{I_{2n+1}}{I_{2n}} \quad (11.2.46)$$

so we need to prove that:

$$\lim_{n \rightarrow \infty} \frac{I_{2n+1}}{I_{2n}} = 1 \quad (11.2.47)$$

Furthermore, for $x \in [0, \pi/2]$ then:

$$\sin^{2n+2} x \leq \sin^{2n+1} x \leq \sin^{2n} x \implies I_{2n+2} \leq I_{2n+1} \leq I_{2n} \quad (11.2.48)$$

using the inequality rules for integrals. Therefore:

$$\frac{I_{2n+2}}{I_{2n}} = \frac{2n+1}{2n+2} \leq \frac{I_{2n+1}}{I_{2n}} \leq 1 \quad (11.2.49)$$

Hence, using the squeeze rule:

$$\lim_{n \rightarrow \infty} \frac{I_{2n+1}}{I_{2n}} = 1 \implies \lim_{n \rightarrow \infty} a_n = \frac{\pi}{2} \quad (11.2.50)$$

as desired.

We then have that:

$$\lim_{n \rightarrow \infty} b_n^2 = \lim_{n \rightarrow \infty} \frac{2n+1}{n} a_n = \pi \implies \lim_{n \rightarrow \infty} b_n = \sqrt{\pi} \quad (11.2.51)$$

■

Series

Theorem (Integral test)

Let f be positive and decreasing on $[1, \infty)$, and suppose $\lim_{x \rightarrow \infty} f(x) = 0$. Define:

$$I_n = \int_1^n f \quad (11.2.52)$$

Then:

(i) $\sum_{n=1}^{\infty} f(n)$ converges if (I_n) is bounded above.

(ii) $\sum_{n=1}^{\infty} f(n)$ diverges if (I_n) diverges.

Proof. Let $s_n = \sum_{k=1}^n f(k)$ be the n th partial sum of $\sum_{n=1}^{\infty} f(n)$, and let P_{n-1} be the partition of $[1, n]$:

$$P_n = \{[1, 2], \dots, [i, i+1], \dots, [n-1, n]\} \quad (11.2.53)$$

Now since $f(x)$ is decreasing on $[0, \infty)$, we must have that:

$$m_i = f(i+1) \implies L(f, P_{n-1}) = f(2) + f(3) + \dots + f(n) = s_n - f(1) \quad (11.2.54)$$

$$M_i = f(i) \implies U(f, P_{n-1}) = f(1) + f(2) + \dots + f(n-1) = s_n - f(n) \quad (11.2.55)$$

Hence:

$$s_n - f(1) \leq \int_1^n f \leq s_n - f(n) \quad (11.2.56)$$

(i) Suppose $I_n = \int_1^n f$ is bounded above by some M :

$$s_n - f(1) \leq I_n \leq M \implies s_n \leq M + f(1) \quad (11.2.57)$$

Since (s_n) is an increasing bounded sequence, it follows from the Monotone convergence theorem that s_n converges. Hence $\sum_{n=1}^{\infty} f(n)$ converges.

(b) Suppose I_n is not bounded above, since f is positive:

$$I_{n+1} - I_n = \int_n^{n+1} f \geq 0 \implies I_n \text{ is increasing} \quad (11.2.58)$$

So we have that I_n diverges. Note also that:

$$s_n \geq I_n \quad (11.2.59)$$

so using the Squeeze rule, s_n diverges, and hence so does $\sum_{n=1}^{\infty} f(n)$. ■

Example. Consider:

$$\int \frac{dx}{x(\log x)^2} = \int \frac{du}{u^2} = -\frac{1}{\log x} \quad (11.2.60)$$

where we used $u = \log x$, $du = \frac{1}{x}$. Hence:

$$I_n = \int_2^n \frac{dx}{x(\log x)^2} = \left[\frac{1}{\log x} \right]_2^n = \frac{1}{\log 2} - \frac{1}{\log n} \quad (11.2.61)$$

Note that $x(\log x)^2$ is positive and increasing on $[2, \infty)$, implying that $f = \frac{1}{x(\log x)^2}$ is positive and decreasing on $[2, \infty)$. Furthermore:

$$I_n = \frac{1}{\log 2} - \frac{1}{\log n} \leq \frac{1}{\log 2} \quad (11.2.62)$$

so I_n is bounded above. It follows that:

$$\sum_{n=2}^{\infty} \frac{1}{n(\log n)^2} \text{ is convergent} \quad (11.2.63) \quad \blacktriangleleft$$

Unit F4: Power series

12.1 Taylor series

Definition (*Taylor polynomial*)

Let $f \in \mathcal{C}^n(I)$ be a n -times differentiable function defined on an open interval I containing a . Then the Taylor polynomial of degree n at a is the polynomial:

$$T_n(x) = \sum_{k=0}^n \frac{f^{(k)}(x)}{k!}(x-a)^k = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \dots + \frac{f^{(n)}(a)}{n!}(x-a)^n \quad (12.1.1)$$

Example. Consider $f(x) = \cos x$, let us find its n th order Taylor polynomial about $a \in \mathbb{R}$. Firstly we evaluate the derivatives:

$$\begin{aligned} f(x) &= \cos x \implies f(a) = \cos a \\ f'(x) &= -\sin x \implies f'(a) = -\sin a \\ f''(x) &= -\cos x \implies f''(a) = -\cos a \\ f^{(3)}(x) &= \sin x \implies f^{(3)}(a) = \sin a \\ f^{(4)}(x) &= \cos x \implies f^{(4)}(a) = \cos a \\ &\vdots \\ f^{(2n)}(x) &= (-1)^n \cos x \implies f^{(2n)}(a) = (-1)^n \cos a \\ f^{(2n+1)}(x) &= (-1)^{n+1} \sin x \implies f^{(2n+1)}(a) = (-1)^{n+1} \sin a \end{aligned}$$

Consequently, even order taylor polynomials are:

$$T_{2n}(a) = \sum_{k=0}^n \frac{(-1)^k \cos a}{(2k)!}(x-a)^{2k} + \sum_{k=0}^{n-1} \frac{(-1)^{k+1} \sin a}{(2k+1)!}(x-a)^{2k+1} \quad (12.1.2)$$

while odd order taylor polynomials are:

$$T_{2n+1}(a) = \sum_{k=0}^n \frac{(-1)^k \cos a}{(2k)!} (x-a)^{2k} + \sum_{k=0}^n \frac{(-1)^{k+1} \sin a}{(2k+1)!} (x-a)^{2k+1} \quad (12.1.3)$$

Taking $a = 0$ we find that:

$$T_{2n}(0) = T_{2n+1}(0) = \sum_{k=0}^n \frac{(-1)^k}{(2k)!} (x-a)^{2k} \quad (12.1.4)$$



Example. Consider $f(x) = \sin x$, let us find its n th order Taylor polynomial about $a \in \mathbb{R}$. Firstly we evaluate the derivatives:

$$\begin{aligned} f(x) &= \sin x \implies f(a) = \sin a \\ f'(x) &= \cos x \implies f'(a) = \cos a \\ f''(x) &= -\sin x \implies f''(a) = -\sin a \\ f^{(3)}(x) &= -\cos x \implies f^{(3)}(a) = -\cos a \\ f^{(4)}(x) &= \sin x \implies f^{(4)}(a) = \sin a \\ &\vdots \\ f^{(2n)}(x) &= (-1)^n \sin x \implies f^{(2n)}(a) = (-1)^n \sin a \\ f^{(2n+1)}(x) &= (-1)^n \cos x \implies f^{(2n+1)}(a) = (-1)^n \cos a \end{aligned}$$

Consequently, even order taylor polynomials are:

$$T_{2n+2}(a) = \sum_{k=0}^{n+1} \frac{(-1)^k \sin a}{(2k)!} (x-a)^{2k} + \sum_{k=0}^n \frac{(-1)^k \cos a}{(2k+1)!} (x-a)^{2k+1} \quad (12.1.5)$$

while odd order taylor polynomials are:

$$T_{2n+1}(a) = \sum_{k=0}^n \frac{(-1)^k \sin a}{(2k)!} (x-a)^{2k} + \sum_{k=0}^n \frac{(-1)^k \cos a}{(2k+1)!} (x-a)^{2k+1} \quad (12.1.6)$$

Taking $a = 0$ we find that:

$$T_{2n+1}(0) = T_{2n+2}(0) = \sum_{k=0}^n \frac{(-1)^k}{(2k+1)!} (x-a)^{2k+1} \quad (12.1.7)$$



Example. Consider $f(x) = e^x$, let us find its n th order Taylor polynomial about $a \in \mathbb{R}$. First we evaluate the derivatives:

$$f^{(n)}(a) = e^a \quad (12.1.8)$$

so that:

$$T_n(a) = \sum_{k=1}^n \frac{e^a}{k!} (x-a)^k \implies T_n(0) = \sum_{k=1}^n \frac{x^k}{k!} \quad (12.1.9)$$



Theorem (Taylor's theorem)

Let $f \in \mathcal{C}^{n+1}(I)$ and $a, x \in I$. Then:

$$f(x) = T_n(x) + R_n(x) \quad (12.1.10)$$

where $T_n(x)$ is the n th order Taylor about a , and:

$$R_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!} (x-a)^{n+1} \quad (12.1.11)$$

for some $c \in (a, x)$ is known as the error term.

Proof. Let's consider:

$$h(t) = f(t) - T_n(t) - A(t-a)^{n+1} \quad (12.1.12)$$

where A is chosen so that $h(x) = 0$. Note also that:

$$f^{(k)}(a) = T_n^{(k)}(a) \implies h^{(k)}(a) = 0, \quad k = 0, 1, \dots, n \quad (12.1.13)$$

from which it follows that h is continuous and n -fold differentiable on I , and that $h^{(k)}(a) = h^{(k)}(x) = 0$. Using Rolle's theorem applied to h on the interval $[a, x]$, we see that there must be some c_1 such that $h'(c_1) = 0$.

Similarly, applying Rolle's theorem to h' on the interval $[a, c_1]$, we see that there must be some c_2 such that $h''(c_2) = 0$. Repeating this reasoning to $h'', h^{(3)}, \dots, h^{(n)}$ on the intervals:

$$[a, c_2], [a, c_3], \dots, [a, c_n], \quad c_2 > c_3 > \dots > c_n > a \quad (12.1.14)$$

we find that there must be some $c \in [a, c_n]$ such that:

$$h^{(n+1)}(c) = f^{(n+1)}(c) - A(n+1)! = 0 \implies A = \frac{f^{(n+1)}(c)}{(n+1)!} \quad (12.1.15)$$

Substituting this into our original expression for $h(t)$, and setting $t = x$ with $h(x) = 0$:

$$f(x) = T_n(x) + \frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1} \quad (12.1.16)$$

as desired.

Example. The Taylor expansion of $f(x) = \log x$ about $a = 0$ is:

$$T_n(x) = \sum_{k=1}^n \frac{(-1)^{k+1} x^k}{k} \quad (12.1.17)$$

for $x \in (-1, 1]$. Let us limit the domain to $I = [-0.02, 0.02] = [a-r, a+r]$ where $r = 0.02$. The second order polynomial is therefore:

$$T_2(x) = x - \frac{x^2}{2}, \quad x \in I \quad (12.1.18)$$

implying that:

$$|R_2(x)| = \left| \frac{f^{(3)}(c)}{3!} x^3 \right| \leq |f^{(3)}(c)| \frac{r^3}{3!} \quad (12.1.19)$$

for some $c \in I$. Now we have that:

$$|f^{(3)}(c)| = \left| \frac{2}{(1+c)^3} \right| \leq 2 \quad (12.1.20)$$

so that:

$$|R_2(x)| \leq 2 \cdot \frac{(0.02)^3}{3!} = 2.66 \times 10^{-6} \quad (12.1.21)$$

so we know that the error will be negligible up to 5 decimal places. Consequently:

$$\log(1.02) \approx (0.02) - \frac{(0.02)^2}{2} \approx 0.01980 \text{ (5 d.p.)} \quad (12.1.22)$$



Example. The fourth order Taylor polynomial $T_4(x)$ at π for $f(x) = \cos x$ is:

$$T_4(x) = \sum_{k=0}^4 \frac{(-1)^{k+1}}{(2k)!} (x-\pi)^{2k} = -1 + \frac{1}{2}(x-\pi)^2 \quad (12.1.23)$$

We need to show that $T_4(\pi)$ approximates $f(\pi) = \cos \pi$ to at least a 0.01 error on $[3\pi/4, 5\pi/4] = [a-r, a+r]$ where $a = \pi$, $r = \frac{\pi}{4}$.

To do so we must find an upper bound for the remainder $|R_4(x)| = \left| \frac{f^{(5)}(c)}{5!} (x-a)^5 \right|$.

Now $|x - a| \leq r = \frac{\pi}{4}$, so:

$$|R_4(x)| \leq |f^{(5)}(c)| \frac{r^5}{5!} \quad (12.1.24)$$

for some $c \in [3\pi/4, 5\pi/4]$. We also have that:

$$f^{(5)}(x) = -\sin x \implies |f^{(5)}(c)| = |\sin c| \leq 1 \quad (12.1.25)$$

and thus:

$$|R_4(x)| \leq \frac{(\pi/4)^5}{5!} = 0.0025 < 0.01 \quad (12.1.26)$$

as desired. ◀

Theorem (Taylor series)

Let f be a class \mathcal{C}^∞ on an open interval I at points a and x . If $R_n(x) \rightarrow 0$ as $n \rightarrow \infty$ then

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x - a)^n \quad (12.1.27)$$

which is known as the **Taylor series** at a for f .

Proof. This follows immediately from the fact that $f(x) = T_n(x) + R_n(x)$. ■

Proposition (Important taylor series at 0)

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n, \quad |x| < 1 \quad (12.1.28)$$

$$\sin x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!}, \quad x \in \mathbb{R} \quad (12.1.29)$$

$$\cos x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!}, \quad x \in \mathbb{R} \quad (12.1.30)$$

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}, \quad x \in \mathbb{R} \quad (12.1.31)$$

$$\log(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1} x^n}{n}, \quad -1 < x \leq 1 \quad (12.1.32)$$

Proof. (a) It can easily be seen that the n th order taylor polynomial of $f(x) = \frac{1}{1-x}$ is:

$$T_n(x) = \sum_{k=0}^n x^k \quad (12.1.33)$$

which is a geometric series. It converges to $\frac{1}{1-x}$ only for $|x| \leq 1$, as desired.

- (b) The n th order taylor polynomial of $f(x) = \sin x$ is:

$$T_n(x) \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!} \quad (12.1.34)$$

implying that the error term $R_n(x)$ may be expressed as:

$$|R_n(x)| = \left| \frac{f^{(n+1)}(c)}{(n+1)!} x^{n+1} \right| \leq \frac{x^{n+1}}{(n+1)!} \quad (12.1.35)$$

since $f^{(n+1)}(c) = \pm \sin c$ or $\pm \cos c$. Therefore:

$$|R_n(x)| \leq \frac{|x|^{n+1}}{(n+1)!} \rightarrow 0 \text{ as } n \rightarrow \infty \quad (12.1.36)$$

where we have used the squeeze rule for null sequences. This is true for all x , and the result follows.

- (c) The n th order taylor polynomial of $f(x) = \cos x$ is:

$$T_n(x) \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!} \quad (12.1.37)$$

implying that the error term $R_n(x)$ may be expressed as:

$$|R_n(x)| = \left| \frac{f^{(n+1)}(c)}{(n+1)!} x^{n+1} \right| \leq \frac{x^{n+1}}{(n+1)!} \quad (12.1.38)$$

since $f^{(n+1)}(c) = \pm \sin c$ or $\pm \cos c$. Therefore:

$$|R_n(x)| \leq \frac{|x|^{n+1}}{(n+1)!} \rightarrow 0 \text{ as } n \rightarrow \infty \quad (12.1.39)$$

where we have used the squeeze rule for null sequences. This is true for all x , and the result follows.

- (d) The n th order taylor polynomial of $f(x) = e^x$ is:

$$T_n(x) = \sum_{k=0}^n \frac{x^k}{k!} \quad (12.1.40)$$

The error term may be written as:

$$R_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!} x^{n+1} = e^c \frac{x^{n+1}}{(n+1)!} \quad (12.1.41)$$

and since c lies between 0 and x :

$$|R_n(x)| \leq e^x \frac{|x^{n+1}|}{(n+1)!} \rightarrow 0 \text{ as } n \rightarrow \infty \quad (12.1.42)$$

as desired.

(e) The n th order Taylor polynomial of $f(x) = \log(1+x)$ is:

$$T_n(x) = \sum_{k=1}^n \frac{(-1)^{k+1} x^k}{k} \quad (12.1.43)$$

Consequently, we have that for $0 < x \leq 1$ then the error term reads:

$$|R_n(x)| = \left| \frac{f^{(n+1)}(c)}{(n+1)!} x^{n+1} \right| \quad (12.1.44)$$

$$= \left| \frac{n!}{(1+c)^{n+1}} \frac{x^{n+1}}{(n+1)!} \right| \quad (12.1.45)$$

$$= \frac{|x|^{n+1}}{(n+1)(c+1)^{n+1}} \leq \frac{1}{n} \rightarrow 0 \text{ as } n \rightarrow \infty \quad (12.1.46)$$

where we noted that $|x|^{n+1} \leq 1$ and $1+c > 1$. Using the squeeze rule we readily find the desired result. ■

12.2 Convergence

Definition (Power series) Let $a, a_n, x \in \mathbb{R}$ for $n = 0, 1, 2, \dots$. Then a power series at a in x is a series of the form:

$$\sum_{n=0}^{\infty} a_n (x-a)^n \quad (12.2.1)$$

Lemma. If the power series $\sum_{n=0}^{\infty} a_n x^n$ converges for some $x_0 \neq 0$, then it converges absolutely on $(-|x_0|, |x_0|)$.

Proof. Let $r = |x_0|$. Note that the convergence of $\sum_{n=0}^{\infty} a_n x_0^n$ implies that $(a_n x_0^n)$ is a null sequence, and hence there exists some K such that:

$$|a_n|r^n = |a_n x_0^n| \leq K, \quad n = 0, 1, 2, \dots \quad (12.2.2)$$

Let $|x| < r$, then clearly:

$$\sum_{n=0}^{\infty} a_n x^n = \sum_{n=0}^{\infty} a_n r^n \frac{x^n}{r^n} \implies |a_n x^n| \leq K \frac{|x|^n}{r^n} \quad (12.2.3)$$

Since we assumed that $|x| < r$, we have that the geometric series $\sum_{n=0}^{\infty} \left(\frac{|x|}{r}\right)^n$ is convergent. By the comparison test, it follows that $\sum_{n=0}^{\infty} a_n x^n$ converges for $|x| < |x_0|$. ■

Theorem (Radius of convergence)

The power series $\sum_{n=0}^{\infty} a_n(x-a)^n$ exactly one of the following is true:

- (a) it converges only for $x = a$
- (b) it converges for all x
- (c) there exists some $R > 0$ such that the series diverges if $|x - a| > R$ and converges if $|x - a| < R$.

and in all cases absolute convergence follows on the same intervals.

Proof. Let us define:

$$E = \left\{ x \in \mathbb{R} : \sum_{n=0}^{\infty} a_n(x-a)^n \text{ converges} \right\} \quad (12.2.4)$$

If $E = \{a\}$ then we have satisfied condition (a).

If E is unbounded, then for all $x \in \mathbb{R}$ there is some $x_0 \in E$ obeying $|x| < |x_0|$. It follows that $\sum_{n=0}^{\infty} a_n(x-a)^n$ converges absolutely for $(-|x_0|, |x_0|)$ by the Lemma we have just proven. Since this holds for all $x \in \mathbb{R}$ the power series must satisfy condition (b).

The only remaining set is bounded and containing some $x_0 \neq a$. Consequently the power series converges absolutely over $(-|x_0|, |x_0|) \subseteq E$. We see that the radius of convergence is $R = \sup E$, so that $R > |x_0|$.

In the case where $|x - a| < R$, then there exists some $x_1 \in E$ such that $|x - a| < x_1$. Therefore, the series converges absolutely.

In the case where $|x - a| > R$, then we can find $x_2 > R$ such that $|x - a| > x_2$. If $\sum_{n=0}^{\infty} a_n(x-a)^n$ were to converge then we would find that $\sum_{n=0}^{\infty} a_n x_2^n$ converges, a contradiction.

So here condition (c) is satisfied. ■

Theorem (Ratio test)

Suppose that $\sum_{n=0}^{\infty} a_n(x-a)^n$ is a power series with radius of convergence R .

- (a) If $\left|\frac{a_{n+1}}{a_n}\right| \rightarrow \infty$ as $n \rightarrow \infty$ then $R = 0$.
- (b) If $\left|\frac{a_{n+1}}{a_n}\right| \rightarrow 0$ as $n \rightarrow \infty$ then $R = \infty$.
- (c) If $\left|\frac{a_{n+1}}{a_n}\right| \rightarrow L$ as $n \rightarrow \infty$ then $R = \frac{1}{L}$ provided $L > 0$.

Proof. (a) Suppose that $\left|\frac{a_{n+1}}{a_n}\right| \rightarrow \infty$ as $n \rightarrow \infty$. If $x \neq a$ then:

$$\frac{|a_{n+1}(x-a)^{n+1}|}{|a_n(x-a)^n|} = \left|\frac{a_{n+1}}{a_n}\right| |x-a| \rightarrow \infty \text{ as } n \rightarrow \infty \quad (12.2.5)$$

proving that $\sum_{n=0}^{\infty} |a_n(x-a)^n|$ diverges. Since absolute convergence of power series follows from normal convergence, we have that $\sum_{n=0}^{\infty} a_n(x-a)^n$ diverges. So the series only converges when $x = a$, giving $R = 0$.

(b) Suppose that $\left|\frac{a_{n+1}}{a_n}\right| \rightarrow 0$ as $n \rightarrow \infty$. If $x \neq a$ then:

$$\frac{|a_{n+1}(x-a)^{n+1}|}{|a_n(x-a)^n|} = \left|\frac{a_{n+1}}{a_n}\right| |x-a| \rightarrow 0 \text{ as } n \rightarrow \infty \quad (12.2.6)$$

proving that $\sum_{n=0}^{\infty} |a_n(x-a)^n|$ converges. Since absolute convergence of power series follows from normal convergence, we have that $\sum_{n=0}^{\infty} a_n(x-a)^n$ diverges. So the series only converges when $x = a$, giving $R = 0$.

(c) Suppose that $\left|\frac{a_{n+1}}{a_n}\right| \rightarrow L$ as $n \rightarrow \infty$. If $x \neq a$ then:

$$\frac{|a_{n+1}(x-a)^{n+1}|}{|a_n(x-a)^n|} = \left|\frac{a_{n+1}}{a_n}\right| |x-a| \rightarrow L|x-a| \text{ as } n \rightarrow \infty \quad (12.2.7)$$

If $|x-a| > \frac{1}{L}$ then we find that:

$$\frac{|a_{n+1}(x-a)^{n+1}|}{|a_n(x-a)^n|} \rightarrow L|x-a| > 1 \text{ as } n \rightarrow \infty \quad (12.2.8)$$

proving that $\sum_{n=0}^{\infty} |a_n(x-a)^n|$ diverges over this interval, and hence so does $\sum_{n=0}^{\infty} a_n(x-a)^n$. It follows that $R \leq \frac{1}{L}$. If $|x-a| < \frac{1}{L}$, then we find that:

$$\frac{|a_{n+1}(x-a)^{n+1}|}{|a_n(x-a)^n|} \rightarrow L|x-a| < 1 \text{ as } n \rightarrow \infty \quad (12.2.9)$$

proving that $\sum_{n=0}^{\infty} |a_n(x-a)^n|$ converges over this interval, and hence so does $\sum_{n=0}^{\infty} a_n(x-a)^n$. It follows that $R \geq \frac{1}{L}$.

Together, these results show that $R = \frac{1}{L}$ as desired. ■

Example. Consider the power series (about 0):

$$\sum_{n=1}^{\infty} \frac{(n!)^2}{(2n)!} x^n \quad (12.2.10)$$

We see that $a_n = \frac{(n!)^2}{(2n)!}$, and hence:

$$\left| \frac{a_{n+1}}{a_n} \right| = \frac{((n+1)!)^2}{(n!)^2} \quad (12.2.11)$$

$$\frac{(2n)!}{(2n+2)!} = \frac{(n+1)^2}{(2n+1)(2n+2)} \quad (12.2.12)$$

$$= \frac{n+1}{2(2n+1)} \rightarrow \frac{1}{4} \text{ as } n \rightarrow \infty \quad (12.2.13)$$

Using the ratio test for power series we see that $R = 4$.

Consider the power series (about 0):

$$\sum_{n=1}^{\infty} n^n x^n \quad (12.2.14)$$

We see that $a_n = n^n$, and hence:

$$\left| \frac{a_{n+1}}{a_n} \right| = \frac{(n+1)^{n+1}}{n^n} = \left(1 + \frac{1}{n} \right)^n (n+1) \rightarrow \infty \text{ as } n \rightarrow \infty \quad (12.2.15)$$

Using the ratio test for power series we see that $R = 0$.

Consider the power series (about 0):

$$\sum_{n=1}^{\infty} (n + 2^{-n})(x - 1)^n \quad (12.2.16)$$

We see that $a_n = (n + 2^{-n})$, and hence:

$$\left| \frac{a_{n+1}}{a_n} \right| = \frac{n+1+2^{-n-1}}{n+2^{-n}} \quad (12.2.17)$$

$$1 + \frac{2^{-n-1} - 2^{-n}}{n+2^{-n}} + \frac{1}{n+2^{-n}} \quad (12.2.18)$$

$$= 1 + \frac{1}{2(n2^n + 1)} + \frac{1}{n+2^{-n}} \rightarrow 1 \text{ as } n \rightarrow \infty \quad (12.2.19)$$

Using the ratio test for power series we see that $R = 1$. ◀

Example. Let's consider the following power series:

$$\sum_{n=0}^{\infty} \frac{\alpha(\alpha-1)\dots(\alpha-n+1)}{n!} x^n \quad (12.2.20)$$

where α is not an integer. Then this is a power series about 0 with coefficients:

$$a_n = \frac{\alpha(\alpha - 1)\dots(\alpha - n + 1)}{n!} \quad (12.2.21)$$

We see that:

$$\left| \frac{a_{n+1}}{a_n} \right| = \left| \frac{\alpha(\alpha - 1)\dots(\alpha - n)}{\alpha(\alpha - 1)\dots(\alpha - n - 1)} \frac{n!}{(n+1)!} \right| \quad (12.2.22)$$

$$= \left| \frac{\alpha - n}{n + 1} \right| \rightarrow 1 \text{ as } n \rightarrow \infty \quad (12.2.23)$$



12.3 The combination rules

Proposition (Combination rules)

Let f, g be a function represented by a Taylor series at a :

$$f(x) = \sum_{n=0}^{\infty} a_n (x - a)^n, \quad |x - a| < R \quad (12.3.1)$$

$$g(x) = \sum_{n=0}^{\infty} b_n (x - a)^n, \quad |x - a| < R' \quad (12.3.2)$$

then for $r = \min\{R, R'\}$ and $\lambda \in \mathbb{R}$:

$$(f + g)(x) = \sum_{n=0}^{\infty} (a_n + b_n) (x - a)^n, \quad |x - a| < r \quad (12.3.3)$$

$$\lambda f(x) = \sum_{n=0}^{\infty} (\lambda a_n) (x - a)^n, \quad |x - a| < R \quad (12.3.4)$$

$$(12.3.5)$$

Note that the theorem does not state that the radius of convergence is $r = \min\{R, R'\}$, it may be larger.

Example. Let us find the taylor series at 0 for $f(x) = \cosh x$. We use the identity:

$$f(x) = \cosh x = \frac{e^x + e^{-x}}{2} = \frac{1}{2} \sum_{n=0}^{\infty} (1 + (-1)^n) \frac{x^n}{n} = \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!} \quad (12.3.6)$$

with infinite radius of convergence.



Proposition (Product rule)

Let f, g be a function represented by a Taylor series at a :

$$f(x) = \sum_{n=0}^{\infty} a_n (x-a)^n, \quad |x-a| < R \quad (12.3.7)$$

$$g(x) = \sum_{n=0}^{\infty} b_n (x-a)^n, \quad |x-a| < R' \quad (12.3.8)$$

then for $r = \min\{R, R'\}$:

$$(fg)(x) = \sum_{n=0}^{\infty} c_n (x-a)^n, \quad |x-a| < r \quad (12.3.9)$$

where

$$c_n = \sum_{k=0}^n a_k b_{n-k} \quad (12.3.10)$$

Example. Let's consider the following function:

$$f(x) = (1+x) \log(1+x) \quad (12.3.11)$$

The taylor series for $\log(1+x)$ at 0 is:

$$\log(1+x) = \sum_{n=0}^{\infty} \frac{(-1)^{n+1} x^n}{n}, \quad -1 < x \leq 1 \quad (12.3.12)$$

implying that:

$$(1+x) \log(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} (x^n + x^{n+1}) \quad (12.3.13)$$

$$= \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n + \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^{n+1} \quad (12.3.14)$$

$$= \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n + \sum_{n=2}^{\infty} \frac{(-1)^n}{n-1} x^n \quad (12.3.15)$$

$$= x + \sum_{n=2}^{\infty} \left(\frac{(-1)^{n+1}}{n} + \frac{(-1)^n}{n-1} \right) x^n \quad (12.3.16)$$

$$= x + \sum_{n=2}^{\infty} (-1)^n \frac{x^n}{n(n-1)} \quad (12.3.17)$$

for $-1 < x \leq 1$.

Let's consider the following function:

$$f(x) = \frac{1}{(1-x)^2} \quad (12.3.18)$$

The taylor series for $\frac{1}{1-x}$ is:

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n, |x| < 1 \quad (12.3.19)$$

Therefore:

$$\frac{1}{(1-x)^2} = \sum_{n=0}^{\infty} c_n x^{2n}, |x| < 1 \quad (12.3.20)$$

where

$$c_n = \sum_{k=0}^n 1 = n+1 \implies \frac{1}{(1-x)^2} = \sum_{n=0}^{\infty} (n+1)x^n, |x| < 1 \quad (12.3.21)$$

Consider the function:

$$f(x) = \frac{1}{1+2x^2} \quad (12.3.22)$$

The taylor series for $\frac{1}{1+x}$ is:

$$\frac{1}{1+x} = \sum_{n=0}^{\infty} (-1)^n x^n, |x| < 1 \quad (12.3.23)$$

Consequently:

$$\frac{1}{1+2x^2} = \sum_{n=0}^{\infty} (-1)^n (2x^2)^n = \sum_{n=0}^{\infty} (-2)^n x^{2n}, |2x^2| < 1 \quad (12.3.24)$$

so the range of validity for this expansion is $2x^2 < 1 \implies |x| \leq \frac{1}{\sqrt{2}}$.

Consider the function

$$f(x) = \frac{e^x}{(1-x)^2} \quad (12.3.25)$$

The taylor series reads:

$$\sum_{n=0}^{\infty} c_n x^n, |x| < 1 \quad (12.3.26)$$

where

$$c_n = \sum_{k=0}^n \frac{k+1}{(n-k)!} \quad (12.3.27)$$

so:

$$c_0 = 1, c_1 = 1+2=3, c_2 = \frac{1}{2} + 2 + 3 = \frac{11}{2} \quad (12.3.28)$$



Theorem (Differentiation rule) The following taylor series:

$$f(x) = \sum_{n=0}^{\infty} a_n(x-a)^n, \quad (12.3.29)$$

$$g(x) = \sum_{n=1}^{\infty} n a_n (x-a)^{n-1} \quad (12.3.30)$$

have the same radius of convergence, and $f'(x) = g(x)$ for $|x-a| < R$.

Theorem (Integration rule) The following taylor series:

$$f(x) = \sum_{n=0}^{\infty} a_n(x-a)^n, \quad (12.3.31)$$

$$F(x) = \sum_{n=0}^{\infty} \frac{a_n}{n+1} (x-a)^{n+1} \quad (12.3.32)$$

have the same radius of convergence R , and if $R > 0$ then:

$$\int f(x) dx = F(x), \quad |x-a| < R \quad (12.3.33)$$

Proof. The two series have the same radius of convergence by applying the differentiation rule to $F(x)$. The differentiation rule also implies that $F'(x) = f(x)$ over $|x-a| < R$, giving the desired integral. ■

Example. Let's find the taylor series at 0 for $f(x) = \tanh^{-1} x$. We have that:

$$f'(x) = \frac{1}{1-x^2} = \sum_{n=0}^{\infty} x^{2n}, \quad |x| < 1 \quad (12.3.34)$$

Consequently:

$$f(x) = \sum_{n=0}^{\infty} \frac{x^{2n+1}}{2n+1}, \quad |x| < 1 \quad (12.3.35)$$

Let's find the taylor series of e^{-x^2} :

$$e^{-x^2} = \sum_{n=0}^{\infty} \frac{(-x^2)^n}{n!} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} x^{2n}, \quad x \in \mathbb{R} \quad (12.3.36)$$

implying that:

$$\int_0^1 e^{-x^2} = \sum_{n=0}^{\infty} \int_0^1 (-1)^n n! x^{2n} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \frac{1}{2n+1} \quad (12.3.37)$$

We define the error function as:

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{n!} \frac{1}{2n+1} \Rightarrow \int_0^1 e^{-x^2} = \frac{\sqrt{\pi}}{2} \operatorname{erf}(1) \quad (12.3.38)$$

Finally, let's find the taylor series of $f(x) = \log(1+x)$. We know that:

$$\frac{1}{1+x} = \sum_{n=0}^{\infty} (-1)^n x^n, \quad |x| < 1 \quad (12.3.39)$$

implying that:

$$\log(1+x) = \int \frac{1}{1+x} dx \sum_{n=0}^{\infty} \frac{(-1)^n}{n+1} x^{n+1} = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n \quad (12.3.40)$$

for $|x| < 1$. ◀

Theorem (General binomial theorem) Let $\alpha \in \mathbb{R}$, then:

$$(1+x)^\alpha = \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n, \quad |x| < 1 \quad (12.3.41)$$

Proof. Let us define:

$$f(x) = \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n, \quad g(x) = f(x)(1+x)^{-\alpha} \quad (12.3.42)$$

Differentiating g we find that:

$$g'(x) = f'(x)(1+x)^{-\alpha} - \alpha f(x)(1+x)^{-\alpha-1} \quad (12.3.43)$$

$$= (1+x)^{-\alpha-1} ((1+x)f'(x) - \alpha f(x)) \quad (12.3.44)$$

$$= (1+x)^{-\alpha-1} \left((1+x) \sum_{n=1}^{\infty} \binom{\alpha}{n} nx^{n-1} - \alpha \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n \right) \quad (12.3.45)$$

We can simplify the expression in brackets:

$$(1+x) \sum_{n=1}^{\infty} \binom{\alpha}{n} nx^{n-1} - \alpha \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n \quad (12.3.46)$$

$$= \sum_{n=1}^{\infty} \binom{\alpha}{n} nx^{n-1} + \sum_{n=1}^{\infty} \binom{\alpha}{n} nx^n - \alpha \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n \quad (12.3.47)$$

$$= \sum_{n=0}^{\infty} \binom{\alpha}{n+1} (n+1)x^n + \sum_{n=1}^{\infty} \binom{\alpha}{n} (n-\alpha)x^n \quad (12.3.48)$$

We find that:

$$\binom{\alpha}{n+1} (n+1) = \frac{\alpha!}{(n+1)!(\alpha-n-1)!} (n+1) = \frac{\alpha!}{(n)!(\alpha-n-1)!} \quad (12.3.49)$$

and

$$\binom{\alpha}{n} (n-\alpha) = \frac{\alpha!}{n!(\alpha-n)!} (n-\alpha) = -\frac{\alpha!}{n!(\alpha-n-1)!} \quad (12.3.50)$$

implying that:

$$\binom{\alpha}{n+1} (n+1) + \binom{\alpha}{n} (n-\alpha) = \frac{\alpha!}{(n)!(\alpha-n-1)!} - \frac{\alpha!}{n!(\alpha-n-1)!} = 0 \quad (12.3.51)$$

Consequently, we see that $g'(x) = 0$, that is, $g(x) = f(x)(1+x)^{-\alpha}$ takes a constant value. Evaluating $g(0)$ we get the desired result. ■

Part II

Algebra and Group Theory

Unit B1: Symmetry and groups

13.1 Symmetry in \mathbb{R}^2

Symmetries of plane figures

A special class of transformations of the plane are called **isometries**. They are transformations that preserve distances between points, and include **reflections**, **rotations**, **translation**, and **glide reflections** (reflections followed by a translation parallel to line of reflection).

Symmetries are special isometries, that maps a bounded plane figure to itself.

Definition 14.1 (Symmetry of plane figures)

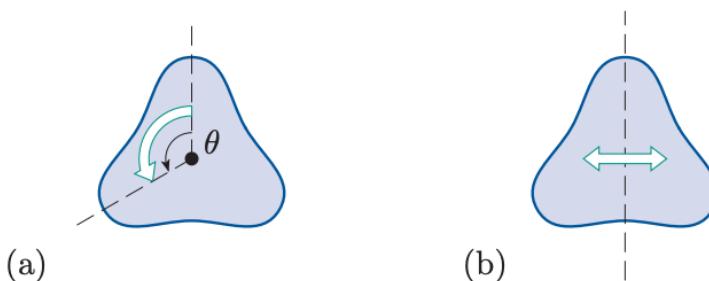
A symmetry of a plane figure \mathcal{F} is an isometry:

$$f : \mathbb{R}^2 \rightarrow \mathbb{R}^2 \quad (13.1.1)$$

$$\mathcal{F} \mapsto \mathcal{F} \quad (13.1.2)$$

For bounded plane figures, translations do not map the figure to itself, and so neither do glide-reflections. They are not symmetries, leaving only rotations and reflections. We also have the **identity transformation**, which leaves every point in \mathbb{R}^2 as they are.

It is important to note that when specifying the angle through which a rotation occurs, this angle must be measured anti-clockwise by convention.



The axes of symmetry of a figure all pass through a point, called the centre of the bounded figure, is also the center of all rotational symmetries.

Consider the symmetries of a square. To keep track of its orientation and position we mark a dot on its upper left corner, and color its other side darker (this allows us to distinguish a rotation by $\pi/2$ from a reflection about a horizontal axis of symmetry for example).

We then find that the square has four rotational symmetries including the identity symmetry) as shown below.

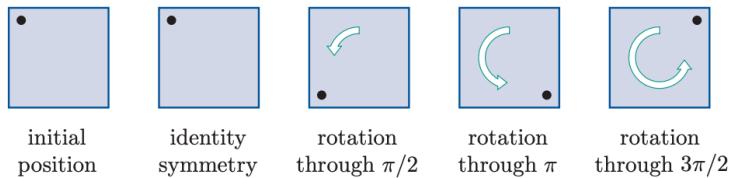


Figure 13.1. Rotational symmetry of a square

The square also has 4 reflection symmetries, 1 horizontal, 1 vertical and 2 diagonal:

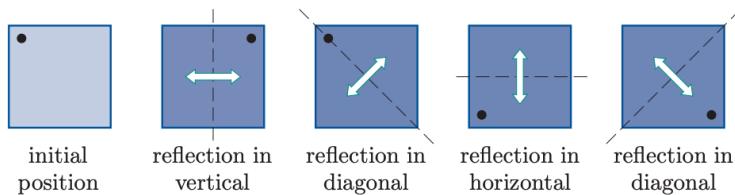


Figure 13.2. Reflection symmetries of a square

In general, it can be shown that a **regular polygon**, that is, a bounded plane figure with n equilateral sides, has $2n$ symmetries.

Theorem 14.2 (*Symmetries of n -gon*)

A regular n -gon has $2n$ symmetries, namely n rotations through $\frac{2\pi}{k}$, $k \in \mathbb{N}^*$ and n reflections. The set of all symmetries is called D_{2n} , and is called the **dihedral group**.

We shall go back to defining the dihedral group more rigorously in chapter 17.

Identities and subsets of $S(\mathcal{F})$

Proposition 14.3 (*Properties of $S(\mathcal{F})$*) The set of symmetries $S(\mathcal{F})$ of a bounded plane figure \mathcal{F} satisfies the following properties

- (i) **Closure under composition:** if $f, g \in S(\mathcal{F})$ then $g \circ f \in S(\mathcal{F})$

- (ii) **Associativity:** if $f, g, h \in S(\mathcal{F})$ then $h \circ (g \circ f) = (h \circ g) \circ f$
- (iii) **Identity existence:** for each symmetry $f \in S(\mathcal{F})$, $f \circ e = e \circ f = f$ where e is the identity symmetry.
- (iv) **Inverse existence:** for each symmetry $f \in S(\mathcal{F})$, $\exists f^{-1} \in S(\mathcal{F})$ such that $f \circ f^{-1} = f^{-1} \circ f = e$, where e is the identity symmetry.

Consider for example the symmetries of a square labelled below:

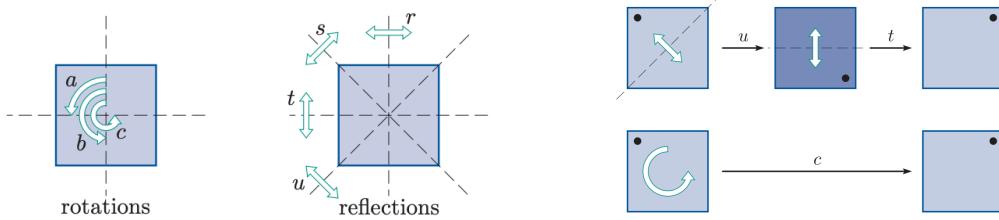


Figure 13.3. Symmetries of a square

Then clearly, if we apply $t \circ u$ as shown below, it is equivalent to applying c , so $t \circ u = c \in S(\mathcal{F})$.

More generally we can write that:

\circ	rotation	reflection
rotation	rotation	reflection
reflection	reflection	rotation

It is also interesting to note that the composition of the same reflection twice gives the identity symmetry i.e. $r \circ r = s \circ s = t \circ t = u \circ u = e$. We say that the reflection symmetries are **self-inverse**.

The inverses of the square symmetries can be summarised as:

Element	e	a	b	c	r	s	t	u
Inverse	e	c	b	a	r	s	t	u

Definition 14.4 (Direct and Indirect symmetries)

Direct symmetries are symmetries of a plane figure \mathcal{F} that do not require us to lift the figure out of \mathbb{R}^2 and flip it. The set of direct symmetries of \mathcal{F} is denoted as $S^+(\mathcal{F})$.

Indirect symmetries are symmetries of a plane figure \mathcal{F} that require us to lift the figure out of \mathbb{R}^2 and flip it.

For bounded plane figures it is immediate that rotations are direct symmetries and reflections are indirect symmetries. Therefore, for a square:

$$S^+(\square) = \{e, a, b, c\} \quad (13.1.3)$$

We then find that:

\circ	direct	indirect
direct	direct	indirect
indirect	indirect	direct

from which it follows that if $f \circ f^{-1} = e$, since e is a direct symmetry, f and its inverse must have the same nature of directness (inverse of direct is direct, inverse of indirect is indirect).

Theorem 14.5 (Number of direct and indirect symmetries)

If a plane figure has finite symmetries, either

- all symmetries are direct
- half the symmetries are direct and the other half indirect

Proof. Consider a plane figure \mathcal{F} with finite symmetries with n direct symmetries. If it has no indirect symmetries, than this case falls under category 1. Consider the case where \mathcal{F} has at least one indirect symmetry. But then, we can compose this indirect symmetry with n direct symmetries creating $n - 1$ other indirect symmetries (since e is unique and one of the n direct symmetries, its composite with the indirect symmetry gives the same transformation, so we don't count it). So the figure also has n indirect symmetries. This algorithm is shown for $S(\square)$: ■

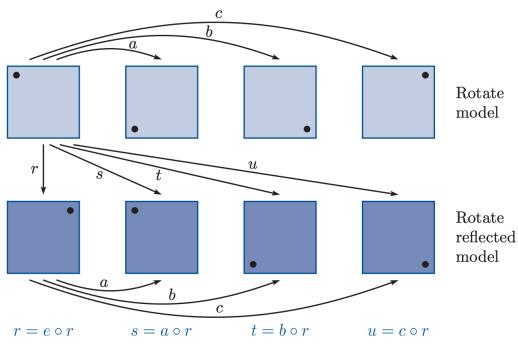


Figure 13.4. Deriving indirect symmetries from direct symmetries

An immediate consequence of this theorem is that no bounded plane figure can have solely indirect symmetry. This is easy to see applying the closure property. Indeed, if f, g are indirect symmetries then $f \circ g$ must be a direct symmetry.

13.2 Representing symmetries

Two line symbol

Because labelling each transformation by a letter may be time consuming and impractical for figures with several symmetries, we introduce the notation of **two line symbols**.

Consider for example the transformation r of the square with the vertices labelled as shown:
We can then represent this transformation as:

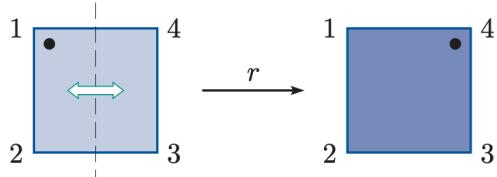


Figure 13.5. Two line symmetry notation

$$r \leftrightarrow \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix} \quad (13.2.1)$$

where the top row shows the initial vertex, and the second row shows where it gets mapped.

Definition 14.6 (Two line symbol)

The two line symbol representing a symmetry f of a polygon \mathcal{F} with vertices $1, 2, 3 \dots n$ is:

$$f \leftrightarrow \begin{pmatrix} 1 & 2 & 3 & \dots & n \\ f(1) & f(2) & f(3) & \dots & f(n) \end{pmatrix} \quad (13.2.2)$$

The inverse f^{-1} is then clearly represented by:

$$f^{-1} \leftrightarrow \begin{pmatrix} f(1) & f(2) & f(3) & \dots & f(n) \\ 1 & 2 & 3 & \dots & n \end{pmatrix} \quad (13.2.3)$$

Cayley tables

The tables we have used so far to categorize composites of reflections and rotations, direct and indirect symmetries are called **Cayley tables**. They can be constructed by listing all the elements of $S(\mathcal{F})$ on the top and left hand side of a square array. For any $x, y \in S(\mathcal{F})$, their composite $x \circ y$ is in the x th row and y th column.

The Cayley table for the symmetries of a rectangle is shown below:

Since we chose to list the direct symmetries and indirect symmetries separately, we formed four blocks each containing only direct or indirect symmetries.

	e	a	r	s			
e	e	a	r	s			
a	a	e	s	r			
r	r	s	e	a			
s	s	r	a	e			

Table 13.1. Rectangle symmetries and its Cayley table

The same occurs with the Cayley table for a square:

	e	a	b	c	r	s	t	u	
e	e	a	b	c	r	s	t	u	
a	a	b	c	e	s	t	u	r	
b	b	c	e	a	t	u	r	s	
c	c	e	a	b	u	r	s	t	
r	r	u	t	s	e	c	b	a	
s	s	r	u	t	a	e	c	b	
t	t	s	r	u	b	a	e	c	
u	u	t	s	r	c	b	a	e	

	direct	indirect	
direct	direct	indirect	
indirect	indirect	direct	

Figure 13.6. Block patterns in Cayley table

13.3 Definition of a Group

Definition 14.7 (Binary operation)

A **binary operation** * is a transformation mapping two members of a set G to another member of G :

$$*: G \times G \longrightarrow G \quad (13.3.1)$$

$$(f, g) \longmapsto h \quad (13.3.2)$$

where $f, g, h \in G$. G is therefore **closed** under *.

If we combine the set G with the binary operation *, then we get a mathematical structure known as a **group**.

Definition 14.8 (Group)

Let G be a set and * be a binary operation on G . Then, $(G, *)$ is a **group** provided $\forall f, g, h \in G$

- (i) **Associativity:** $f * (g * h) = (f * g) * h$
- (ii) **Identity existence:** $\exists e \in G$, called **identity element** such that $f * e =$

$$e * f = f$$

- (iii) **Inverse existence:** $\exists g^{-1} \in G$ called the **inverse** of g such that $g * g^{-1} = g^{-1} * g = e$

Example. Let $X = \{(a, b) \in \mathbb{R}^2 : a \neq 0\}$ and $(a, b) * (c, d) = (ac, ad + b)$. Show that $(X, *)$ is a group.

- (i) **Closure:** let $(a, b), (c, d) \in X$, so $a, b, c, d \in \mathbb{R}$ with $a \neq 0$ and $c \neq 0$. Then:

$$(a, b) * (c, d) = (ac, ad + b) \in \mathbb{R} \quad (13.3.3)$$

since $ac, ad + b \in \mathbb{R}$ and $ac \neq 0$ because $a \neq 0$ and $c \neq 0$.

- (ii) **Associativity:** let $(a, b), (c, d), (e, f) \in X$ then:

$$((a, b) * (c, d)) * (e, f) = (ac, ad + b) * (e, f) = (ace, acf + ad + b) \quad (13.3.4)$$

and

$$(a, b) * ((c, d) * (e, f)) = (a, b) * (ce, cf + d) = (ace, acf + ad + b) \quad (13.3.5)$$

The two expressions are equivalent, as required, so associativity is satisfied.

- (iii) **Identity:** let (e_1, e_2) be the identity element of the group. Then, we need $\forall (a, b) \in X$:

$$(a, b) * (e_1, e_2) = (ae_1, ae_2 + b) = (a, b) \quad (13.3.6)$$

$$(e_1, e_2) * (a, b) = (ae_1, be_1 + e_2) = (a, b) \quad (13.3.7)$$

which upon equating terms component-wise gives:

$$\begin{cases} ae_1 = a \\ ae_2 + b = b \\ be_1 + e_2 = b \end{cases} \implies \begin{cases} e_1 = 1 \\ e_2 = 0 \end{cases} \quad (13.3.8)$$

So the identity element is $(1, 0)$. Let us prove this. Firstly, $(1, 0) \in X$ since $1 \neq 0$ and it belongs to \mathbb{R}^2 . Then, $\forall (a, b) \in X$:

$$(a, b) * (1, 0) = (a \cdot 1, a \cdot 0 + b) = (a, b) \quad (13.3.9)$$

$$(1, 0) * (a, b) = (1 \cdot a, 1 \cdot b + 0) = (a, b) \quad (13.3.10)$$

as required, $(1, 0)$ is the identity element of $*$ over X .

(iv) Consider $(a, b) \in X$, and suppose (c, d) is its inverse. Therefore:

$$(a, b) * (c, d) = (ac, ad + b) = (1, 0) \quad (13.3.11)$$

$$(c, d) * (a, b) = (ac, bc + d) = (1, 0) \quad (13.3.12)$$

which implies:

$$\begin{cases} ac = 1 \\ ad + b = 0 \\ bc + d = 0 \end{cases} \implies \begin{cases} c = \frac{1}{a} \\ d = -\frac{b}{a} \end{cases} \quad (13.3.13)$$

where we used $a \neq 0$ by definition since $(a, b) \in X$. Therefore the inverse of (a, b) is $(\frac{1}{a}, -\frac{b}{a})$. To prove this, consider that $(\frac{1}{a}, -\frac{b}{a}) \in X$ because $\frac{1}{a}, \frac{b}{a} \in \mathbb{R}$ and $\frac{1}{a} \neq 0$ for $a \neq 0$. Also:

$$(a, b) * \left(\frac{1}{a}, -\frac{b}{a}\right) = \left(a \frac{1}{a}, -\frac{b}{a}a + b\right) = (1, 0) \quad (13.3.14)$$

$$(c, d) * \left(\frac{1}{a}, -\frac{b}{a}\right) = \left(a \frac{1}{a}, b \frac{1}{a} - \frac{b}{a}\right) = (1, 0) \quad (13.3.15)$$

as required. So the inverse of (a, b) is $(\frac{1}{a}, -\frac{b}{a})$, and every element in X has an inverse.

Since all group axioms are satisfied, $(X, *)$ is a group. ◀

Note that the operation $*$ need not to be commutative. In such cases we refer to the group as **Abelian**.

Definition 14.9 (Abelian group)

A group $(G, *)$ where $*$ is commutative, that is, $\forall f, g \in G, f \circ g = g \circ f$, is called an **Abelian group**.

Definition 14.10 (Finite and Infinite groups)

A group $(G, *)$ is said to be **finite** if $|G| = n < \infty$.

A group is said to be **infinite** if G is an infinite set.

In Unit A1 we saw that a field $(F, +, \times)$ is a field provided it satisfies the twelve field axioms. With our newfound knowledge of groups, we may now provide an alternative definition:

1. $(F, +)$ is an Abelian group
2. $(F \setminus \{0\}, \times)$ is an Abelian group
3. the distributive law holds for all $x, y, z \in F$, so $x \times (y + z) = x \times y + x \times z$

Checking group axioms with Cayley tables

Checking the group axioms 1 for a finite set can be done using a Cayley table.

Consider for example the group $(\mathbb{Z}_4, +_4)$, whose Cayley table is:

$+_4$	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

We can clearly see that every element in the body of the table belongs to \mathbb{Z}_4 , so \mathbb{Z}_4 is closed under $+_4$, and the latter is a binary operation. We also know from module A2 that $+_4$ is associative, so the first group axiom is satisfied. The identity element is 0, because the row and column labelled 0 repeat the borders of the table. Finally, because each row and column contains the identity element 0, this means that every element of \mathbb{Z}_4 has an inverse. Indeed, $0^{-1} = 0$, $1^{-1} = 3$, $2^{-1} = 2$, $3^{-1} = 1$.

Hence $(\mathbb{Z}_4, +_4)$ satisfies the group axioms, and is therefore a group.

Note that just showing that one column labelled 0 or one row labelled 0 repeats the borders of the table is not enough to show that 0 is the identity element. Both row and column must repeat the borders.

Proposition 14.11 (Reading e from Cayley tables)

Let $(G, *)$ be a group, then e is the identity element of the group iff both the group and column labelled e repeat the table borders.

Proof. In the Cayley table, the row and column labelled e contains all elements $e * g$ and $g * e$ respectively. So, saying that the row/column labelled e repeats the borders of the table is equivalent to saying that $e * g = g$ and $g * e = g$ for all $g \in G$. So e is the identity element of G . ■

\circ	... a b c ...
⋮	⋮
e	... a b c ...
⋮	⋮
	... e o a e o b e o c ...

\circ	... e ...
⋮	⋮
a	a
b	b
c	c
⋮	⋮

Figure 13.7. Row and column of the identity element e

Proposition 14.12 (Reading inverses from Cayley tables) Let $(G, *)$ be a group, then h is an inverse of g iff e appears both in the position with row g , column h and in the position with row h , column g .

Proof. The element in position with row g , column h is $g * h$ and similarly the element in position with row h , column g is $h * g$. Therefore, claiming that e is in both these positions is equivalent to saying $g * h = h * g = e$, so that h is the inverse of g as required. ■

We saw that $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ are fields under addition and multiplication so that:

$$(\mathbb{Z}, +), (\mathbb{Q}, +), (\mathbb{R}, +), (\mathbb{C}, +), (\mathbb{Q}^*, \times), (\mathbb{R}^*, \times), (\mathbb{C}^*, \times) \quad (13.3.16)$$

are all groups.

Theorem 14.13 (\mathbb{Z}_n and \mathbb{U}_n)

For $n \geq 2$, the set \mathbb{Z}_n is a group under $+_n$, and the set \mathbb{U}_n of integers in \mathbb{Z}_n coprime to n is a group under \times_n .

Proof. The group axioms for $(\mathbb{Z}_n, +_n)$ hold because they are the properties of addition in \mathbb{Z}_n as explained in Unit 2.

Let us now prove that (\mathbb{U}_n, \times_n) is a group.

Closure

We need to prove that $\forall a, b \in \mathbb{U}_n$, $a \times_n b \in \mathbb{U}_n$. To do so, we use the result from Unit A2 that $a \times_n b$ is co-prime to n provided that it has a multiplicative inverse in \mathbb{Z}_n . If we denote the inverses of a, b as c, d respectively then we can write using the commutativity of \times_n :

$$(c \times_n d) \times_n (a \times_n b) = (c \times_n a) \times_n (d \times_n b) = 1 \times_n 1 = 1 \quad (13.3.17)$$

and similarly:

$$(a \times_n b) \times_n (c \times_n d) = 1 \quad (13.3.18)$$

so $c \times_n d$ is the multiplicative inverse of $a \times_n b$, and the latter is therefore co-prime to n .

Associativity

We know that modular multiplication is associative.

Identity

Consider $1 \in \mathbb{U}_n$ and $\forall a \in \mathbb{Z}_n$:

$$a \times_n 1 = 1 \times_n a = a \quad (13.3.19)$$

so 1 is the identity element.

Inverses

Let $a \in \mathbb{U}_n$, which implies that $\exists b, a \times_n b = 1$. We have to show that $b \in \mathbb{U}_n$. To do so, consider:

$$a \times_n b = b \times_N a = 1 \quad (13.3.20)$$

so that b is also co-prime to n , and therefore $b \in \mathbb{U}_n$. Hence a has an inverse in \mathbb{U}_n .

Hence (\mathbb{U}_n, \times_n) satisfies all group axioms and is therefore a group. ■

An immediate consequence of this theorem is that $(\mathbb{Z}_p^*, \times_p)$ is a group provided p is prime.

13.4 Properties of groups and group elements

Proposition 14.14 (Properties of groups)

For a group $(G, *)$ the following properties hold:

- (i) the identity element e is unique i.e. $g * e = g, g * e' = g \implies e = e'$
- (ii) each element has a unique inverse i.e. $g * h = e, g * h' = e \implies h = h'$
- (iii) the inverse of g^{-1} is g
- (iv) $(g * h)^{-1} = h^{-1} * g^{-1}$
- (v) if $g * h = g * f$ then $h = f$ (left cancellation law) and if $h * g = f * g$ then $h = f$ (right cancellation law)
- (vi) $g^m * g^n = g^{m+n}$ for $m, n \in \mathbb{Z}$
- (vii) $(g^m)^n = g^{mn}$ for $m, n \in \mathbb{Z}$.

Proof.

- (i) Let $e, e' \in (G, *)$ be both identity elements. Then $e * e' = e$ since e' is an identity element. Similarly, $e * e' = e'$ since e is an identity element. Therefore $e = e'$, and thus the identity element is unique.
- (ii) Let $g * h = e$ and $g * h' = e$, then we may write $h' * g * h = (h' * g) * h = e * h = h$ and similarly $h' * g * h = h' * (g * h) = h' * e = h'$ so that $h = h'$ as required.
- (iii) $g * g^{-1} = g^{-1} * g = e$ implies that g is the inverse of g^{-1} .
- (iv) We first show that $(g * h) * (h^{-1} * g^{-1}) = e$:

$$(g * h) * (h^{-1} * g^{-1}) = g * (h * h^{-1}) * g^{-1} \quad (13.4.1)$$

$$= g * e * g^{-1} \quad (13.4.2)$$

$$= g * g^{-1} \quad (13.4.3)$$

$$= e \quad (13.4.4)$$

We now show that $(h^{-1} * g^{-1}) * (g * h) = e$:

$$(h^{-1} * g^{-1}) * (g * h) = h^{-1} * (g^{-1} * g) * h \quad (13.4.5)$$

$$= h^{-1} * e * h \quad (13.4.6)$$

$$= h^{-1} * h \quad (13.4.7)$$

$$= e \quad (13.4.8)$$

as required.

(v) To prove the left cancellation law, $g * h = g * f \implies g^{-1} * g * h = g^{-1} * g * f \implies (g^{-1} * g) * h = (g^{-1} * g) * f \implies e * h = e * f \implies h = f$ as required. To prove the right cancellation law $h * g = f * g \implies h * g * g^{-1} = f * g * g^{-1} \implies h * (g * g^{-1}) = f * (g * g^{-1}) \implies h * e = f * e \implies h = f$ as required.

$$(vi) g^m * g^n = \underbrace{g * g * \dots * g}_{m \text{ times}} * \underbrace{g * g * \dots * g}_{n \text{ times}} = \underbrace{g * g * g * \dots * g * g}_{m+n \text{ times}} = g^{m+n}$$

$$(vii) (g^m)^n = \underbrace{x * x * \dots * x}_{m \text{ times}} * \underbrace{\dots * x * x * \dots * x}_{m \text{ times}} = \underbrace{g * g * g * \dots * g * g}_{mn \text{ times}} = g^{mn}$$

■

13.5 Symmetry in \mathbb{R}^3

We now adapt the study of symmetry in \mathbb{R}^2 to bounded figures in \mathbb{R}^3 , called **solids**. More specifically we will consider **convex polyhedra**, solids whose faces are polygons, and which have no dents or dimples, nor spikes.

Definition 14.15 (Symmetry)

A symmetry of a figure \mathcal{F} is an isometry:

$$f : \mathbb{R}^3 \rightarrow \mathbb{R}^3 \quad (13.5.1)$$

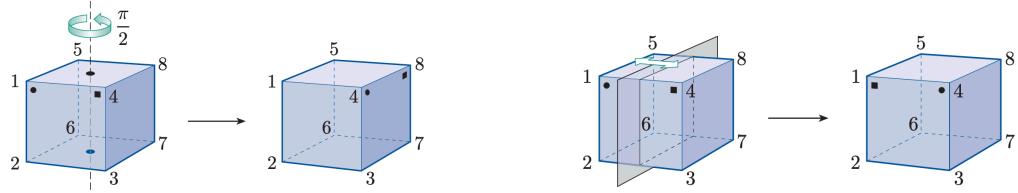
$$\mathcal{F} \mapsto \mathcal{F} \quad (13.5.2)$$

Two symmetries f, g are **equal** if $f(X) = g(X), \forall X \in \mathcal{F}$

The potential symmetries of a bounded plane figure in \mathbb{R}^3 are:

1. identity transformation
2. rotation specified by an axis of symmetry, direction and angle of rotation
3. reflection in a plane
4. composite of the above

The two line symbol applies as always. For example, consider the rotation of a cube through $\pi/2$ about its vertical axis:



- (a) Rotation of a cube through $\pi/2$ about its vertical axis (b) Reflection of a cube in the vertical plane

Figure 13.8.

Using two-line symbols we can write:

$$g \leftrightarrow \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 3 & 7 & 8 & 1 & 2 & 6 & 5 \end{pmatrix} \quad (13.5.3)$$

Similarly a reflection in the vertical plane is represented by:

$$f \leftrightarrow \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 3 & 2 & 1 & 8 & 7 & 6 & 5 \end{pmatrix} \quad (13.5.4)$$

The composition of the two is performed as intuition would expect, that is, reading off $y = g(x)$ from 14.5.3 and then $f(y)$ from 14.5.5 for all x vertices.

$$f \circ g \leftrightarrow \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 1 & 2 & 6 & 5 & 4 & 3 & 7 & 8 \end{pmatrix} \quad (13.5.5)$$

which is a reflection in the diagonal plane as shown below:

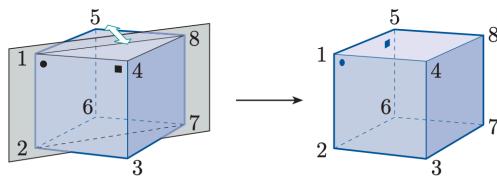


Figure 13.9. Reflection in diagonal plane of a cube

Theorem 14.16 (Symmetry group)

$S(\mathcal{F})$ forms a group under the composition function for bounded figures in \mathbb{R}^n . The group $(S(\mathcal{F}), \circ)$ is called the **symmetry group**.

Analogously to plane figures, the symmetries of solid figures can also be classified as direct or indirect. In this case however, it must be noted that one cannot physically demonstrate

a reflection, since to do so would require accessing the fourth dimension. Therefore, we may define direct symmetries as those which can be shown physically in \mathbb{R}^3 , and indirect symmetries as all the others.

Finding the number of symmetries of a polyhedron

Theorem 14.17 (*Symmetries of regular polyhedra*)

For a regular polyhedron with F faces each with n symmetries, then the number of symmetries of the polyhedron is $F \cdot n$.

To see why this is the case, consider a tetrahedron, a polyhedron made up of 4 equilateral triangular faces. We wish to find the number of ways of replacing the figure in the space it occupied originally.

We can choose any one of the four equilateral triangles as a base, each with 6 symmetries. For each symmetry of the base, we can allow the entire tetrahedron to be transformed accordingly, resulting in a symmetry of the solid. So there are $4 \times 6 = 24$ symmetries.

For irregular polyhedra, the process is similar. We consider for example a small rhombicuboctahedron, with 18 squares and 8 equilateral triangles as faces.

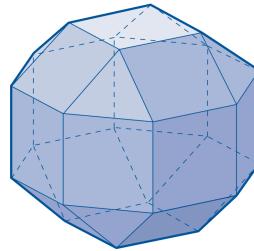


Figure 13.10. Rhombicuboctahedron

Again we look at all the ways of replacing the polyhedron in the space it occupied originally. To do so we first find a type of face we can use as base, such as a square such as the square faces.

Now immediately we realize that not all squares of the rhombicuboctahedron can be used as a base. Indeed, we have two different types of square faces that give different symmetries. We will choose the type of square faces to the left in Figure 14.11 as our base.

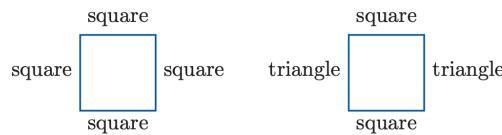


Figure 13.11. Square faces of a rhombicuboctahedron

We then see that only 6 square faces are suitable squares of this type. Each have 8 symmetries which give symmetries of the polyhedron (this is not always the case, for some polyhedra not all symmetries of a base will correspond in symmetries of the figure). Therefore, there are 48 total symmetries.

If we chose the second type of square face (of which there are 12), then only 4 of the symmetries of the square would have been suitable. Indeed, rotations by $\pi/2$ and $3\pi/2$, as well as the two reflections in the diagonals do not give a symmetry of the polyhedron. Hence, as was found earlier, there are 48 symmetries.

Finding the symmetries of a polyhedron

Let us try to find all the symmetries of a regular tetrahedron. Using Theorem 14.6 one can easily show that it has 12 symmetries. We start by finding all direct symmetries. As always we have the identity symmetry:

$$e = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix} \quad (13.5.6)$$

For each base, there is a rotational symmetry about a fixed axis through the opposite vertex.

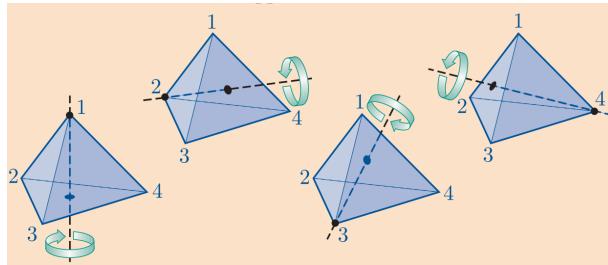


Figure 13.12. Rotational symmetries of the tetrahedron about axes through bases

One can write the two line symbol for each of these symmetries, getting to a total of 9 direct symmetries. We are therefore missing three, which can be found by composing direct symmetries with each other.

Having found all direct symmetries, we now try to find one indirect symmetry. For example, a reflection in the vertical plane as shown:

We can then compose this reflection symmetry with the 12 direct symmetries to find 12 indirect symmetries. This accounts for all 24 symmetries of the tetrahedron.

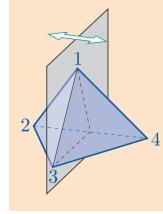


Figure 13.13. Reflectional symmetry of tetrahedron

13.6 The Dihedral group

Theorem 14.18 (Order of D_{2n})

The **Dihedral group** D_{2n} , the group of symmetries of a regular n -gon, has order $2n$.

Proof. We will consider the polygon $\mathcal{F} \subseteq \mathbb{C}$, with vertices at $e^{2im\pi} n, 0 \leq m \leq n$.

We define the following map:

$$r : \mathbb{C} \longrightarrow \mathbb{C} \quad (13.6.1)$$

$$z \mapsto z \cdot e^{\frac{2i\pi}{n}} \quad (13.6.2)$$

which is a rotation about the center of the polygon by $\frac{2\pi}{n}$. This is a symmetry, since it preserves distances $\forall z, w \in \mathbb{C}$:

$$|r(z) - r(w)| = |(z - w) \cdot e^{\frac{2i\pi}{n}}| = |z - w| \quad (13.6.3)$$

We define the reflection in the x -axis by the following map:

$$s : \mathbb{C} \longrightarrow \mathbb{C} \quad (13.6.4)$$

$$z \mapsto \bar{z} \quad (13.6.5)$$

which is again another symmetry since $\forall z, w \in \mathbb{C}$:

$$|s(z) - s(w)| = |\bar{z} - \bar{w}| = \sqrt{(\bar{z} - \bar{w})(z - w)} = |z - w| \quad (13.6.6)$$

We will show that:

$$D_{2n} = \underbrace{\{e, r, r^2, r^3, \dots, r^{n-1}\}}_{\text{rotations}}, \underbrace{\{s, rs, r^2s, \dots, r^{n-1}s\}}_{\text{reflections}} \quad (13.6.7)$$

so that all symmetries of \mathcal{F} are some composition of r and s .

Indeed, let $f \in D_{2n}$ so that $1 \in \mathcal{F} \implies f(1) \in \mathcal{F}$, 1 gets mapped to another vertex, say $e^{2\pi ik}n$ for some $0 \leq k < n$. However, r^k also maps $1 \mapsto e^{2\pi ik}n$ so that $g \equiv r^{-k} \circ f$ is an isometry fixing 1.

Lemma. The composite of two isometries over the metric space (\mathbb{C}, d) where $d(z, w) = |z - w|$, is an isometry over the same metric space.

Proof. Let f, g be two such isometries. Then:

$$d((f \circ g)(z), (f \circ g)(w)) = |(f \circ g)(z) - (f \circ g)(w)| \quad (13.6.8)$$

$$= |f(g(z)) - f(g(w))| \quad (13.6.9)$$

$$= |g(z) - g(w)| = |z - w| \quad (13.6.10)$$

$$= d(z, w) \quad (13.6.11)$$

so $f \circ g$ is an isometry over (\mathbb{C}, d) . ■

Since $e^{2\pi i}n$ shares an edge with 1, and since g preserves distances, we require $g(e^{2\pi i}n)$ to also share an edge with 1.

The two possibilities are either $e^{2\pi i}n$ or $e^{2\pi i(n-1)}n$.

In the first case where g fixes 1 and $e^{2\pi i}n$. We can repeat the same argument as before for $g(e^{4\pi i}n)$, which can only get mapped to itself in order to preserve distances. Suppose all vertices $e^{2\pi ik}n$ with $k \leq m-1$ for some $0 < m < n$ have been mapped to themselves by g . Then, $e^{2\pi im}n$ can only be mapped to itself or $e^{2\pi i(m-2)}n$. However, the latter cannot be the case, since $g(e^{2\pi i(m-2)}n) = e^{2\pi i(m-2)}n$, the vertex has already been "taken". Consequently, $g(e^{2\pi im}n) = e^{2\pi i(m-2)}n$, and by the principle of induction all vertices have been fixed. Hence, $g = e$, the identity transformation, implying $f = r^k$.

If instead g fixes 1, and $g(e^{2\pi i}n) = e^{2\pi i(n-1)}n$. We then have $(s \circ g)(e^{2\pi i}n) = e^{\frac{2\pi i}{n}}$. Also, $(s \circ g)(1) = s(1) = 1$, hence $s \circ g$ fixes 1 and $e^{\frac{2\pi i}{n}}$. By the same argument as before then, $s \circ g = e$, and consequently $f = r^k \circ s$.

We have therefore proven that any isometry f can be expressed as either r^k , a rotation, or $r^k \circ s$, a reflection. In total, there are n such rotations and n such reflections¹, giving $|D_{2n}| = 2n$ as desired. ■

Proposition 14.19 (Properties of D_{2n})

Let $r, s \in D_{2n}$ be a rotation and reflection respectively, as defined in the previous

¹they are also all distinct. Clearly, if $r^m = r^k$, then 1 gets mapped to $e^{\frac{2\pi im}{n}}$ and $e^{\frac{2\pi ik}{n}}$, giving $m = k$. In other words all the rotations r^i are distinct for all $0 \leq i < n$. Also, $s \neq r^i$ for any i , since s fixes 1, whereas r^i only does so if $r^i = e = s$, which is a contradiction. Finally, $r^i s \neq r^k s$ follows immediately by composing by s to the left

proof. Then:

- (i) $sr^k s = r^{-k}$ for all $0 \leq k < n$
- (ii) $\text{ord}(r) = n$, $\text{ord}(r^i s) = 2$, for all $0 \leq i < n$
- (iii) D_{2n} is not abelian if $n \geq 3$.
- (iv) any rotation is the composition of two reflections

Proof.

- (i) Consider where $r^k s$ maps the vertex $e^{\frac{2\pi im}{n}}$. s sends it to $e^{\frac{-2\pi im}{n}}$, followed by r^k which sends it to $e^{\frac{2\pi i(k-m)}{n}}$. Similarly, sr^{-k} maps $e^{\frac{2\pi im}{n}}$ to $e^{\frac{2\pi i(m-k)}{n}}$, and then to $e^{\frac{2\pi i(k-m)}{n}}$. Hence $r^k s$ and sr^{-k} map $e^{\frac{2\pi im}{n}}$ to the same vertex, and are equivalent. $sr^{-k} = r^k s \implies r^{-k} = sr^k s$.
- (ii) Since $e, r, r^2, \dots, r^{n-1}$ are all distinct, and $r^n = e$, it follows that $\text{ord}(r) = n$, since for any $i < n$, $r^i \neq r^n = e$. Note that by definition, $s^2(z) = s(\bar{z}) = \bar{\bar{z}} = z \implies s^2 = e$. Also, $(r^i s)(r^i s) = (r^i s)(sr^{-i}) = r^i s^2 r^{-i} = r^i r^{-i} = e$, so $\text{ord}(r^i s) = 2$.
- (iii) We have that $rs = sr^{-1}$, and suppose $rs = sr$, then $sr^{-1} = sr \implies r^2 = e$, which is a contradiction since $\text{ord}(r) = n > 2$ by assumption.
- (iv) Consider the rotation r^i for some $0 \leq i < n$. Then, it may be written as $r^i = (r^{i+1}s) \circ (sr)$

■

Unit B2: Subgroups and isomorphisms

14.1 Subgroups

Definition 15.1 (Subgroup)

A **subgroup** of a group $(G, *)$ is a group $(H, *)$ where $H \subseteq G$. We write that $(H, *) \leq (G, *)$.

For example, the group $S^+(\mathcal{F}, \circ)$ is a subgroup of $S(\mathcal{F}, \circ)$ because $S^+(\mathcal{F}) \subseteq S(\mathcal{F})$.

Every non-empty group $(G, *)$ has at least two subgroups, $(e, *)$ called the **trivial subgroup** and $(G, *)$ itself. All subgroups other than $(H, *)$ with $H \subsetneq H$ are called **proper subgroups**.

Theorem 15.2 (Identity and inverses of subgroups)

Let $(G, *)$ be a group with subgroup $(H, *)$. Then:

- (i) the identity element of $(H, *)$ is the same as the identity element of $(G, *)$
- (ii) $\forall h \in H$, the inverse of h in $(G, *)$ and $(H, *)$ is the same.

Proof.

- (i) Let the identity element in $(G, *)$ and $(H, *)$ be e and e_H respectively. We must therefore have $e_H \circ e = e_H$ since e is the identity element of $(G, *)$ and $e_H \circ e = e$ since e_H is the identity element of $(H, *)$. It follows immediately that $e = e_H$.
- (ii) Let the inverses of h in $(H, *)$ be x and y . We then have that $h * a = h * b = e$ where e is the identity element of $(G, *)$ and thus of $(H, *)$. Using the left cancellation law, $a = b$ as required.

■

The astute reader may have noted that some of the group axioms hold for any subgroup of a group. It is therefore only necessary to prove some of the properties of a group to ascertain that it is a subgroup.

Theorem 15.3 (Subgroup criteria)

Let $(G, *)$ be a group with identity element e and let $H \subseteq G$. Then $(H, *)$ is a subgroup of $(G, *)$ if and only if $\forall x, y \in H$

$$(SG1) \quad x * y \in H$$

$$(SG2) \quad e_G \in H$$

$$(SG3) \quad x_G^{-1} \in H$$

where x_G^{-1} is the inverse of x in G and e_G is the identity element of $(G, *)$.

Proof. We begin by proving the \implies implication. Suppose that $(H, *)$ is a subgroup of $(G, *)$, so that $H \subseteq G$. Then, the closure group axiom of H under $*$ asserts that $x * y \in H$. Similarly, the existence of an identity element e_G was proven in Theorem 15.2 (a). Finally the existence of the inverse was proven in Theorem 15.2(b)

We now prove the \impliedby implication. Suppose that (SG1)-(SG3) are satisfied, we must check that the group axioms are satisfied.

Closure is trivial

Associativity since $(G, *)$ is a group, $\forall x, y, z \in (G, *)$ we have that $x * (y * z) = (x * y) * z$.

Since $H \subseteq G$ it follows that $x, y, z \in (H, *) \implies x, y, z \in (G, H*)$ and therefore associativity holds in H as well.

Identity if $e \in G$ then $x \in H \implies x \in G$ and thus $x * e_G = e_G * x = x$ using the identity group axiom of G . $e_G \in H$ is trivial, since it is equivalent to (SG2).

Inverses if $x \in G$ then $x_G^{-1} \in G$ and $x \in H$. Thus $x * x_G^{-1} = x_G^{-1} * x = e_G$ as required. $x_G^{-1} \in H$ is trivial, since it is equivalent to (SG3). ■

Example. Show that $(\{e, a, b, c\}, \circ)$ is a subgroup of $(S(\square), \circ)$

We have that $\{e, a, b, c\} \subseteq S(\square)$ and \circ is a binary operation.

The Cayley table is:

\circ	e	a	b	c
e	e	a	b	c
a	a	b	c	e
b	b	c	e	a
c	c	e	a	b

Closure is clearly satisfied since all elements in the table are $\{e, a, b, c\}$. The identity element in $(S(\square), \circ)$ is e , which is in $\{e, a, b, c\}$. The elements e, b are self-inverse and a, c are inverses of each other so $(\{e, a, b, c\}, \circ)$ contains all the inverses of its elements. ◀

Proposition 15.4 (Integer Multiples subgroups)

The only subgroups of $(\mathbb{Z}, +)$ are $n\mathbb{Z} = \{kn : k \in \mathbb{Z}\}$ for $n \in \mathbb{N}$.

Proof. It is clear that $n\mathbb{Z}$ are subgroups. Let us prove that they are the only subgroups of $(\mathbb{Z}, +)$.

Let $H \leq (\mathbb{Z}, +)$, then we must have $0 \in H$. If no other elements are included, then $H = 0\mathbb{Z}$. If we instead have other elements, we pick the smallest positive integer n in H . Then $H = n\mathbb{Z}$. Otherwise, if this is not the case, then $\exists a \in H$ such that n doesn't divide a . The division algorithm allows us to write $a = p \cdot n + q \in H$, with $0 < q < n$. By closure $a, p \cdot n \in H$, implying $q \in H$. Yet, $q < n$ is smaller than n , the smallest element of H , giving us a contradiction. So there are no elements of H not divisible by its smallest element. Furthermore, to satisfy closure we must have all multiples of n . So, $H = n\mathbb{Z}$. ■

Example. Show that $(A, *)$ is a subgroup of $(X, *)$ with $A = \{(a, b) \in X : a = 1\}$.

(i) **Closure:** Let $(1, b), (1, d) \in A$ then:

$$(1, b) * (1, d) = (1, d + b) \in A \quad (14.1.1)$$

since the first term is 1 and the second belongs to \mathbb{R} .

(ii) **Identity:** the identity element in X is $(1, 0)$ which belongs to A as required.

(iii) **Inverse:** the inverse of $(1, b) \in A$ in $(X, *)$ is given by $(1, -b)$. This element belongs to A so every element in A has an inverse.

Because these subgroup properties are all satisfied, $(A, *)$ is a subgroup of $(X, *)$. ◀

Subgroup of symmetry groups

We saw that the symmetries of a figure form a symmetry group under the composition function. We also saw how the subset of all direct symmetries of a square forms a group under composition, so that $(S^*(\square), \circ)$ is a subgroup of $(S(\square), \circ)$. This is true for any figure as we shall prove now.

Theorem 15.5 (Direct symmetry subgroup) Let $F \subsetneq \mathbb{R}^2(\mathbb{R}^3)$ be a figure. Then $(S^+(\mathcal{F}), \circ)$ is a subgroup of $(S(\mathcal{F}), \circ)$.

Proof. We have $S^+(\mathcal{F}) \subseteq S(\mathcal{F})$ and \circ is the same binary operation on both sets. We now check the three subgroup properties:

(i) **Closure:** composing any two direct symmetries gives a direct symmetry, so $S^+(\mathcal{F})$ is closed.

- (ii) **Identity:** the identity element of $(S(\mathcal{F}), \circ)$ is e , the identity symmetry, which also belongs to $(S^+(\mathcal{F}), \circ)$ since it is direct.
- (iii) **Inverse:** if f is a direct symmetry, then because e is a direct symmetry f^{-1} must also be direct and hence belong to $S^+(\mathcal{F})$.

So all the subgroup properties are satisfied as required. ■

Another way to produce a subgroup is to modify the figure by adding shaded patterns. For example, consider coloring the square as shown below:

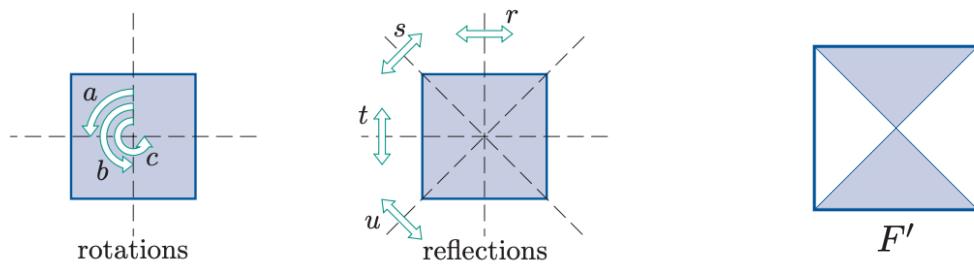


Figure 14.1. Symmetries of \square , and the modified square \mathcal{F}' with symmetries $\{e, b, r, t\}$

Then clearly the symmetries of the figure are restricted. Indeed, we can see that the only symmetries are $S(\mathcal{F}') = \{e, b, r, t\}$, which therefore forms a subgroup under \circ of the symmetry group of a square.

Finally, a third way to find a subgroup of a symmetry group is to fix some feature of the figure (an edge or vertex usually). The resulting symmetries will still form a subgroup, as will be shown now.

Proposition 15.6 (Fixed subset symmetries) Let $\mathcal{F} \subsetneq \mathbb{R}^2(\mathbb{R}^3)$, and let $A \subseteq \mathcal{F}$. Then the subset of $S(\mathcal{F})$ whose elements are all symmetries of \mathcal{F} that fix A is a subgroup under \circ .

Proof. Let H be a subset of $S(\mathcal{F})$ that fixes A . Then:

- (i) **Closure:** if $f, g \in H$, then they both fix A . Hence, if we perform one symmetry after the other A will still remain fixed so $f \circ g \in H$.
- (ii) **Identity:** the identity symmetry fixes A , and so the identity element of $S(\mathcal{F})$ belongs to H .
- (iii) **Inverses:** let $f \in H$, then performing the symmetry in reverse, that is, f^{-1} , A must remain fixed. So H contains the inverse of each of its elements.

The three subgroup properties are satisfied, and thus H forms a subgroup under \circ . ■

Example. Find the elements of the subgroup that consists of all symmetries of the tetrahedron fixing the vertex labelled 4.

The only such symmetries are rotations about lines containing the vertex 4 and reflections about planes containing the vertex 4.

We therefore have that the only direct symmetries are rotations about the line through 4 and the centre of the opposite face, which are three (including e).

There exists an indirect symmetry, a reflection about the plane containing 4, and the height of its opposite face. There must therefore be two other indirect symmetries using Theorem 14.5. Indeed all of the three reflections shown below (they are the only ones containing 4). \blacktriangleleft

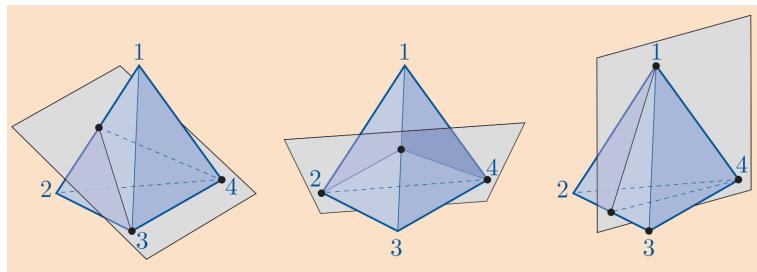


Figure 14.2. Indirect symmetries of the tetrahedron fixing a vertex

We saw in proposition 14.14 that the following properties hold for a group $(G, *)$:

- (i) $g^m * g^n = g^{m+n}$
- (ii) $(g^m)^n = g^{mn}$

Using this index notation to represent the repeated use of a binary operation works if the operation is akin to multiplication. However, it would not work well for operations using addition. For example, in the group $(\mathbb{R}, +)$ one would not denote $x + x + x$ as x^3 but rather $3x$. These two types of notations are known as **multiplicative** and **additive** notation.

We can therefore write:

Proposition 15.7 (Index laws in additive notation)

Let g be an element of a group $(G, *)$ then for $m, n \in \mathbb{Z}$:

- (i) $mx + nx = (m + n)x$
- (ii) $n(mx) = (nm)x$

Definition 15.8 (Group element order)

Let x be an element of a group $(G, *)$. If there exists $n \in \mathbb{N}$ such that $x^n = e$ then the smallest such n is the order of the element x , which has **finite order**.

If there is no such n then we say that x has **infinite order**.

For example the element $a \in S(\square)$ which is the rotation anti-clockwise by $\frac{\pi}{2}$ has order 4. Indeed, $a^4 = e$ is the smallest $n = 4$ displaying this periodicity.

Instead, $2 \in (\mathbb{R}^*, \times)$ has infinite order, because there is no integer n such that $2^n = 1$.

It is also clear that the identity element has order 1. Similarly, any self inverse element $x \neq e$ has order 2.

Proposition 15.9 (Properties of element orders I)

- (i) Let x be an element of a finite group G , then x has finite order.
- (ii) If x is an element of a group, then either x and x^{-1} have the same finite order, or they both have infinite order.

Proof.

- (i) Consider the elements $\dots, x^{-3}, x^{-2}, x^{-1}, x^0, x, x^2, x^3 \dots$ which must belong to G by closure. Because the group is of finite order, at least one element must be repeated, or else we would have infinitely many elements. So, $\exists s, t \in \mathbb{Z}$ with $s < t$ such that:

$$x^s = x^t \implies x^s * (x^s)^{-1} = x^t * (x^s)^{-1} \implies e = x^{t-s} \quad (14.1.2)$$

Since $t - s$ is positive, we have that x has finite order.

- (ii) Let x be an element of a group with identity e . First let us show that $x^n = e \iff (x^{-1})^n = e$. Indeed, suppose that for some $n \in \mathbb{Z}$:

$$x^n = e \implies (x^n)^{-1} * x^n = (x^n)^{-1} \implies (x^{-1})^n = e \quad (14.1.3)$$

so we see that x^{-1} also has the same order. To prove the converse, because the implication has been proven for any element of the group, we simply replace x^{-1} with x . So the values for which $x^n = e$ are the same as the values for which $(x^{-1})^n = e$, so x and x^{-1} have the same order, or both have infinite order.

■

Proposition 15.10 (Properties of element orders II)

Let x be an element of $(G, *)$ then:

- (i) if x has finite order n then:

$$e, x, x^2, \dots, x^{n-1} \quad (14.1.4)$$

are all distinct and repeat every n powers.

- (ii) If x has infinite order, then all powers of x are distinct.

Proof.

- (i) Suppose x has finite order n , and suppose that the powers $e, x, x^2, \dots, x^{n-1}$ are not distinct, so that $x^u = x^t$ for some $0 \leq t < u \leq n - 1$. We can then deduce that $e = x^{u-t}$. However, since $0 < u - t < n$, we see that x has order $u - t$, which is a contradiction. Therefore the above powers of x must all be distinct.

Now consider any integer multiple of n power: x^{kn} with $k \in \mathbb{Z}$. Then we have:

$$x^{kn} = (x^n)^k = e^k = e \quad (14.1.5)$$

since e has order 1. It follows that e repeats every n elements, and since all other powers are formed by composing x , it follows that all the elements listed above also repeat every n elements.

- (ii) Let x have infinite order, so that $\nexists n \in \mathbb{N}$ such that $x^n = e$. Then, suppose that the powers of x are not all distinct, so that for some $0 \leq t < u \leq n - 1$, $x^u = x^t$. But then $x^{u-t} = e$ which implies that x has finite order, contradicting our initial assumption. ■

Example. Find the order of all elements in $(\mathbb{Z}_6, +_6)$.

The identity element 0 has order 1. For the element 1:

$$1 +_6 1 +_6 1 +_6 1 +_6 1 +_6 1 = 0 \quad (14.1.6)$$

so 1 has order 6, and $1^{-1} = 5$ has order 6 as well.

For the element 2:

$$2 +_6 2 +_6 2 = 0 \quad (14.1.7)$$

so 2 has order 3, and $2^{-1} = 4$ has order 3 as well.

Finally the element 3 is self-inverse and therefore has order 2. ◀

14.2 Cyclic groups and subgroups

We can consider the powers of a group element as forming a cycle that repeats itself after n operations.

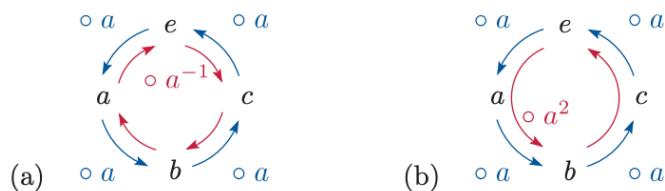


Figure 14.3. Cycle of powers of a in $S(\square)$

We can see in (a) that moving around the cycle anti-clockwise is equivalent to performing a repeatedly. Moving clockwise does the opposite, so it takes a^{-1} . It follows that the element right before the identity element in a cycle of powers of x is x^{-1} .

In (b) we see that moving twice in the clockwise direction is the same as doing $a \circ a = a^2 = b$ and it is then clear that b has order 2.

Definition 15.11 (Generated subset)

Let x be an element of $(G, *)$, then the set of all powers of x is called the **subset of G generated by x** and denoted by $\langle x \rangle = \{x^k : k \in \mathbb{Z}\}$ in multiplicative notation and $\langle x \rangle = \{kx : k \in \mathbb{Z}\}$ in additive notation.

For example, the subset $\langle a \rangle$ of $S(\square)$ consists of the consecutive powers of a in Figure 15.3. Therefore $\langle a \rangle = \{e, a, b, c\}$.

Theorem 15.12 (Cyclic subgroup)

$(\langle x \rangle, *)$ is a subgroup of $(G, *)$ for any element $x \in G$. We call $(\langle x \rangle, *)$ **cyclic subgroup of G generated by x** , and x is a **generator** of $\langle x \rangle$.

Proof.

(i) Closure: let $g, h \in \langle x \rangle$ so that $g = x^s$ and $h = x^t$ for some $s, t \in \mathbb{Z}$. Then:

$$g * h = x^{s+t} \in \langle x \rangle \quad (14.2.1)$$

so we have closure.

(ii) The identity element of $(G, *)$, e , also belongs to $(\langle x \rangle, *)$ since $e = x^0$.

(iii) Let $g \in \langle x \rangle$, then $g = x^s$. Then $g^{-1} = (x^s)^{-1} = x^{-s} \in \langle x \rangle$ so $\langle x \rangle$ includes the inverses of all of its elements.

■

Example. Show that $\langle x \rangle = \langle x^{-1} \rangle$.

Proof. $\langle x \rangle = \{x^k : k \in \mathbb{Z}\} = \{x^{-k} : -k \in \mathbb{Z}\} = \{x^{-k} : k \in -\mathbb{Z}\}$. However, note that $-\mathbb{Z} = \mathbb{Z}$ so that $\langle x \rangle = \{x^{-k} : k \in \mathbb{Z}\} = \langle x^{-1} \rangle$ as required. ■

◀

Since $\langle a \rangle = \{e, a, b, c\}$, and $c = a^{-1}$ we have that $\langle c \rangle = \{e, a, b, c\}$.

Proposition 15.13 (Cyclic subgroup order and element order)

If x has finite order n then the subgroup $(\langle x \rangle, *)$ has order n .

If x has infinite order then so does the subgroup $(\langle x \rangle, *)$.

Definition 15.14 (Cyclic group)

Let $(G, *)$ be a group with element x such that $G = \langle x \rangle$. Then $(G, *)$ is called a **cyclic group**, otherwise, if there is no such x then it is **non-cyclic**.

For example, we saw that $(\mathbb{Z}_6, +_6)$ contains two generators, 1 and 5. Note also that infinite groups can be cyclic, such as $(\mathbb{Z}, +)$ which is generated by both 1 and -1 .

Theorem 15.15 (Abelianity of cyclic (sub)groups)

Every cyclic group is abelian and every subgroup of a cyclic group is cyclic.

Proof. Let us first prove that every cyclic group is abelian. Suppose that $(G, *)$ is a cyclic group with generator a , and let $f, g \in G$, so that $f = x^s$ and $g = x^t$. Then:

$$f * g = a^s * a^t = a^{s+t} = a^{t+s} = a^t * a^s = g * f \quad (14.2.2)$$

so $(G, *)$ is abelian.

Let us now prove that every subgroup $(H, *)$ of $(G, *)$ is cyclic. Suppose $H = \{e\}$, the trivial subgroup, then it is clearly cyclic (generated by e). Suppose now that H is non-trivial, but since $H \subseteq G$ all elements of H are powers of a . Let m be the smallest positive integer such that $a^m \in H$ (it must exist since if $a^m \in H$ then $a^{-m} \in H$ and we must have at least one element in H with non-zero exponent). We will prove that a^m generates H .

Indeed, let $h \in H \implies h = a^k$ for some $k \in \mathbb{Z}$. By the division theorem: $k = qm + r$ for some q, r with $0 \leq r < m$. Then:

$$a^r = a^{k-qm} = a^k * (a^m)^{-q} = h * (a^m)^{-q} \quad (14.2.3)$$

but since H is a group, it must be closed under $*$ and hence $a^r \in H$. But since m is the smallest positive integer such that $a^m \in H$ and $0 \leq r < m$, we must have $r = 0$ to not have a contradiction. Then $k = qm$ and:

$$a^k = (a^m)^q \quad (14.2.4)$$

so we can generate H with a^m as required. Hence $(H, *)$ is cyclic. ■

Example. Find all subgroups of $(\mathbb{Z}_5^*, \times_5)$.

We firstly note that $(\mathbb{Z}_5^*, \times_5)$ is a cyclic group. Indeed looking at the generated

subgroups:

$$\langle 1 \rangle = \{1\} \quad (14.2.5)$$

$$\langle 2 \rangle = \{2, 4, 3, 1\} = \mathbb{Z}_5^* \quad (14.2.6)$$

$$\langle 3 \rangle = \{3, 4, 2, 1\} = \mathbb{Z}_5^* \quad (14.2.7)$$

$$\langle 4 \rangle = \{4, 1\} \quad (14.2.8)$$

we see that $(\mathbb{Z}_5^*, \times_5)$ is generated by 2, and is therefore cyclic. Hence, all the subgroups must be cyclic, and are included in the list above:

$$\{1\}, \{1, 4\}, \mathbb{Z}_5^* \quad (14.2.9)$$



14.3 Cyclic groups and modular arithmetic

We have seen that the additive group $(\mathbb{Z}_6, +_6)$ is cyclic and generated by 1 with order 6. This is true more generally for any group $(\mathbb{Z}_n, +_n)$ with $n \geq 2$, which is cyclic of order n .

Theorem 15.16 (Order of cyclic group elements)

$(\mathbb{Z}_n, +_n)$ is cyclic of order n . Any non-zero element m of the group has order $\frac{n}{d}$ where $d = \text{GCD}(m, n)$.

Proof. We start by proving a useful lemma:

Lemma. Let m be a non-zero element of $(\mathbb{Z}_n, +_n)$, if m is a factor of n then m has order $\frac{n}{m}$.

Indeed, repeatedly adding m in $(\mathbb{Z}_n, +_n)$ is the same as moving m places at a time around the cycle. Starting from 0 then and adding $m \frac{n}{m}$ times we reach 0, so starting from m and adding $m \frac{n}{m}$ times we reach m . Hence the order of m is $\frac{n}{m}$.

Now let m be a non-zero integer in \mathbb{Z}_n and let $d = \text{GCD}(m, n)$. Then $\frac{m}{d}$ and $\frac{n}{d}$ are coprime integers, and by the lemma we have proven, d has order $\frac{n}{d}$.

Our goal is to prove that $\frac{n}{d} \leq \text{ord}(m) \leq \frac{n}{d}$.

To show that $\text{ord}(m) \leq \frac{n}{d}$, consider the cycle of multiples of 1:

If we start from 0 and move around m places at a time $\frac{n}{d}$ times, then we move around a total of $\frac{mn}{d}$. Since $l \equiv \frac{m}{d}$ is an integer, this means that we have gone around l times, and so end up at 0. Hence, $\text{ord}(m)$ is indeed at most $\frac{n}{d}$.

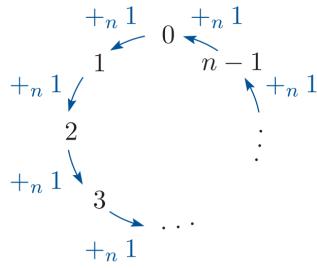


Figure 14.4. Cycle of 1 in $\mathbb{Z}_n, +_n$

Now let us show that $\frac{n}{d} \leq \text{ord}(m)$ by contradiction. Indeed, suppose that $1 \leq \text{ord}(m) = r < \frac{n}{d}$. Then starting from 0 and moving m places at a time r times, we end up at 0 again by definition. Hence we must have for some $k \in \mathbb{N}$:

$$rm = kn \implies r \frac{m}{d} = k \frac{n}{d} \quad (14.3.1)$$

but since $\frac{m}{d}$ and $\frac{n}{d}$ are coprime, the only way the above equation may be true is if $\frac{m}{d}$ is a factor of $k \frac{n}{d}$ and hence $\frac{m}{d}$ is a factor of k . We can then write:

$$rd = \frac{kd}{m} n \quad (14.3.2)$$

where $\frac{kd}{m}$ is an integer since $\frac{m}{d}$ is a factor of k . Thus rd is a multiple of n , and therefore going around the cycle d places at a time r times we end up at 0 again. In other words, $\text{ord}(d) = r = \frac{n}{d}$, which contradicts the assumption $1 \leq \text{ord}(m) = r < \frac{n}{d}$. Thus the order of m cannot be less than $\frac{n}{d}$, and we may conclude that:

$$\text{ord}(m) = \frac{n}{d} \quad (14.3.3)$$

as desired. ■

Corollary 1. For any prime p , any non-zero element m of the group $(\mathbb{Z}_p, +_p)$ has order p .

Proof. Since p is prime, $d = \text{GCD}(m, p) = 1$ and so $\text{ord}(m) = \frac{p}{1} = p$. ■

Corollary 2. A generator m of a group $(\mathbb{Z}_n, +_n)$ must be coprime to n .

Proof. If $m = 0$ then it is not a generator and it also is not coprime to n . Now assume m is non-zero. Then it is a generator iff $\frac{n}{d} = n$ where $d = \text{GCD}(m, n)$. Hence $d = 1$ and thus m, n are coprime as required. ■

This corollary also means that any element of $(\mathbb{Z}_p, +_p)$ is a generator.

Proposition 15.17 (Subgroups of $(\mathbb{Z}_n, +_n)$)

The group $(\mathbb{Z}_n, +_n)$ has exactly one cyclic subgroup of order q for each factor q of n , and no other subgroups.

- (i) the subgroup of order 1 is generated by 0
- (ii) the subgroup of order q is generated by $d = \frac{n}{q}$

Proof. The cyclic subgroup of order 1 is the one generated by 0.

Let q be any factor of n that is not 1 and let $qd = n$. Then, $\text{GCD}(n, d) = d$ and thus d generates a cyclic subgroup $\langle d \rangle$ of order $\frac{n}{d} = q$.

Now let us prove that there are no further cyclic subgroups of order q . Let $m \in \mathbb{Z}_n^*$ and consider $(\langle m \rangle, *)$. If $d = \text{GCD}(n, m)$ then $m \in \langle d \rangle \implies \langle m \rangle \leq \langle d \rangle$ by Theorem 15.12. However, the two subgroups must also have the same order by assumption, so they must be equal. Therefore $\langle q \rangle$ for each factor q are all the subgroups, and they are uniquely determined. ■

14.4 Isomorphisms

Let us compare two cyclic groups of order 4, $(S^+(\square), \circ)$ and $(\mathbb{Z}_4, +_4)$. These two groups are structurally identical.

Indeed, to "move" from one group to another it suffices to $e \leftrightarrow 0$, $a \leftrightarrow 1$, $b \leftrightarrow 2$, $c \leftrightarrow 3$ in their Cayley tables.

\circ	e	a	b	c	$+_4$	0	1	2	3
e	e	a	b	c	0	0	1	2	3
a	a	b	c	e	1	1	2	3	0
b	b	c	e	a	2	2	3	0	1
c	c	e	a	b	3	3	0	1	2

$(S^+(\square), \circ)$ $(\mathbb{Z}_4, +_4)$

Figure 14.5. Cayley tables for $(S^+(\square), \circ)$ and $(\mathbb{Z}_4, +_4)$ and corresponding pattern

Indeed one could entirely remove symbols, and simply color the tiles accordingly to get a pattern. Now consider the group $(\mathbb{Z}_5^*, \times_5)$ which is also of group 4. One can quickly see that its pattern table is:

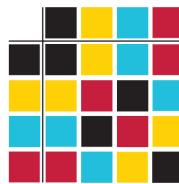


Figure 14.6. Pattern table for $(\mathbb{Z}_5^*, \times_5)$ and its relation with the pattern table for $(\mathbb{Z}_4, +_4)$

The relation with the pattern in Figure 15.5 is then immediate. Indeed, one can switch columns and rows 3,4 to find:

$\times 5$	1	2	3	4		$\times 5$	1	2	4	3		$\times 5$	1	2	4	3
1	1	2	3	4		1	1	2	4	3		1	1	2	4	3
2	2	4	1	3	→	2	2	4	3	1	→	2	2	4	3	1
3	3	1	4	2	swap columns	3	3	1	2	4	swap rows	4	4	3	1	2
4	4	3	2	1	3, 4	4	4	3	1	2	3, 4	3	3	1	2	4
original table				intermediate table				rearranged table								

Because we have to switch columns and rows, we can claim that $(\mathbb{Z}_4, +_4)$ and $(\mathbb{Z}_5^*, \times_5)$ are *not* structurally identical. This concept of "moving" is important and leads to the concept of an isomorphism.

We can define a mapping ϕ between $(S^+(\square), \circ)$ and $(\mathbb{Z}_4, +_4)$ such that:

$$\phi : (S^+(\square), \circ) \longrightarrow \mathbb{Z}_4, +_4 \quad (14.4.1)$$

$$\{e, a, b, c\} \mapsto \{1, 2, 3, 4\} \quad (14.4.2)$$

We see that this type of mapping that transforms one Cayley table to another must be bijective, it must map every element of the first group to exactly one element of the second group.

However, we must have another property, we cannot simply map the elements randomly. Indeed if we used the rule $\{e, a, b, c\} \mapsto \{2, 3, 4, 1\}$ then we would find an entirely different Cayley table shown in the figure below.

\circ	e	a	b	c		2	3	4	1	
e	e	a	b	c	\longrightarrow	2	2	3	4	1
a	a	b	c	e		3	3	4	1	2
b	b	c	e	a		4	4	1	2	3
c	c	e	a	b		1	1	2	3	4

$(S^+(\square), \circ)$

Figure 14.7. Cayley table using $\{e, a, b, c\} \mapsto \{2, 3, 4, 1\}$

This table is clearly not the correct Cayley table for $(\mathbb{Z}_4, +_4)$. Although it has the same structure, the relations between different elements is no longer satisfied (e.g. $1 +_4 3 \neq 2$).

So to have a coherent mapping we must have not only the border elements of the table mapped correctly, but also the body elements. In other words, for any two given elements we must have $\phi(x * y) = \phi(x) * \phi(y)$, $\forall x, y \in G$.

$$\begin{array}{c|ccc}
 \circ & \cdots & y & \cdots \\ \hline
 \vdots & & \vdots & \\
 x & \cdots & x \circ y & \cdots \\ \vdots & & \vdots &
 \end{array} \longrightarrow \begin{array}{c|ccc}
 * & \cdots & \phi(y) & \cdots \\ \hline
 \vdots & & \vdots & \\
 \phi(x) & \cdots & \phi(x \circ y) & \cdots \\ \vdots & & \vdots &
 \end{array}$$

(G, \circ) $(H, *)$

Definition 15.18 (Isomorphic groups)

Two groups $(G, \circ), (H, *)$ are isomorphic if there exists a mapping $\phi : G \rightarrow H$ called **isomorphism** such that:

- (i) ϕ is bijective
- (ii) $\forall x, y \in G \quad \phi(x \circ y) = \phi(x) * \phi(y)$

We then write $(G, \circ) \cong (H, *)$ to assert the isomorphic relation.

We can say that two groups belong to the same **isomorphism class** if they are isomorphic to each other.

Proposition 15.19 (Order of isomorphic groups)

If two groups are isomorphic than they either have both finite order, or they are both infinite.

Example. Let (G, \times) be a cyclic subgroup of (\mathbb{R}, \times) with $G = \{2^k : k \in \mathbb{Z}\}$. Show that:

$$\psi : G \rightarrow \mathbb{Z} \tag{14.4.3}$$

$$2^k \mapsto k \tag{14.4.4}$$

We firstly show that ϕ is bijective, that is, both injective and surjective.

Suppose that $\phi(2^j) = \phi(2^k)$ for $j, k \in \mathbb{Z}$. Then $j = k$ and hence ψ is injective.

Now $\text{Im}(\phi) = \{k \in \mathbb{Z} : 2^k \in G\} = \mathbb{Z}$ and thus surjectivity is satisfied.

Finally, for $2^j, 2^k \in G$:

$$\phi(2^j \times 2^k) = \phi(2^{j+k}) = j + k = \phi(2^j) + \phi(2^k) \tag{14.4.5}$$

so ϕ is indeed an isomorphism. ◀

Proposition 15.20 (Properties of isomorphisms)

Let (G, \circ) and $(H, *)$ be groups with identities e_G, e_H respectively. Then for any isomorphism $\phi : (G, \circ) \rightarrow (H, *)$ and $\forall g \in G$:

- (i) $\phi(e_G) = e_H$

- (ii) $\phi(g^{-1}) = (\phi(g))^{-1}$ (this is actually true for any k in the exponent, but the case $k = -1$ is very important)
- (iii) $\text{ord}(g) = \text{ord}(\phi(g))$
- (iv) if $(K, \circ) \leq (G, \circ)$ then $(\phi(K), *) \leq (H, *)$
- (v) if (G, \circ) is abelian/cyclic then so is $(H, *)$.
- (vi) $\phi(g^k) = (\phi(g))^k, \forall k \in \mathbb{Z}$.

Proof.

- (i) since $e_G \circ e_G = e_G$ we have $\phi(e_G \circ e_G) = \phi(e_G) * \phi(e_G) = \phi(e_G)$. We now rewrite $\phi(e_G) = \phi(e_G) * e_H$ and use the left cancellation law to find that $\psi(e_G) = e_H$.
- (ii) since $g \circ g^{-1} = g^{-1} \circ g = e_G$ we find $\phi(g \circ g^{-1}) = \phi(g^{-1} \circ g) = e_H$ so $\phi(g) * \phi(g^{-1}) = e_H$ thus proving that $\phi(g^{-1}) = (\phi(g))^{-1}$.
- (iii) Suppose $\text{ord}(g) = k$, since ϕ is injective $g^k = e_G \iff \phi(g^k) = (\phi(g))^k = \phi(e_G) = e_H$. Hence $\text{ord}(\phi(g)) \leq \text{ord}(g) = k$. Since ϕ is bijective, it has an inverse ϕ^{-1} , so that repeating the argument of before using $\text{ord}(\phi(g)) = l$ one finds: $\phi^{-1}(\phi(g)^l) = (\phi^{-1}(\phi(g)))^l = g^l = e_G$. Consequently, $l = \text{ord}(\phi(g)) \geq \text{ord}(g)$. Finally, we find that $\text{ord}(\phi(g)) = \text{ord}(g)$.
- (iv) we prove the three subgroup properties for $\phi(K)$:

Closure: let $l_1 = \phi(k_1)$ and $l_2 = \phi(k_2)$ for some $k_1, k_2 \in K$. Then: $l_1 * l_2 = \phi(k_1) * \phi(k_2) = \phi(k_1 \circ k_2) \in \phi(K)$ since $k_1 \circ k_2 \in K$ by the closure property of subgroups.

Identity: $e_H = \phi(e_G) \in \phi(K)$

Inverses: let $l = \phi(k)$ for some $k \in K$. Then $l^{-1} = (\phi(k))^{-1} = \phi(k^{-1}) \in \phi(K)$ since $k^{-1} \in K$ by the subgroup properties of K .

- (v) suppose that (G, \circ) is abelian, and let $h_1 = \phi(g_1), h_2 = \phi(g_2)$ for some $g_1, g_2 \in G$. Then $g_1 \circ g_2 = g_2 \circ g_1 \implies \phi(g_1 \circ g_2) = \phi(g_2 \circ g_1)$ and hence $\phi(g_1) * \phi(g_2) = \phi(g_2) * \phi(g_1)$ and thus $(\phi(K), *)$ is abelian as well.

Now suppose that (G, \circ) is cyclic and generated by a . Let $h = \phi(g)$ for some $g \in G$ then $g = a^k \implies \phi(g) = \phi(a^k) = (\phi(a))^k$ so $(H, *)$ is generated by $\phi(a)$ and is thus cyclic as well.

- (vi) The case for $k = 1$ is trivial. Now, suppose that for some $k \geq 1$, $\phi(g^k) = (\phi(g))^k$. Then, $\phi(g^{k+1}) = \phi(g \circ g^k) = \phi(g) \circ \phi(g^k) = (\phi(g))^{k+1}$. Hence, by the principle of mathematical induction, $\phi(g^k) = (\phi(g))^k, \forall k \in \mathbb{N}$. The case for $k = 0, -1$ have been proven in (i) and (ii) respectively. To prove this statement $\forall k \in \mathbb{Z}$, simply repeat the induction proof, but use g^{-1} instead of g , and apply property (ii). ■

Note that one can use the above identities to prove that two groups are not isomorphic:

1. if one group has more/less number of self-inverse elements
2. if one group has a different order than the other
3. if one group is abelian/cyclic, and the other is not

Proposition 15.21 (*Isomorphisms of cyclic groups*)

Let (G, \circ) and $(H, *)$ be finite cyclic groups of order n or infinite groups generated by a, b respectively. Then they are isomorphic through $\phi : a^k \mapsto b^k$, $k = 0, 1, 2 \dots n-1$ for finite ordered groups and $k \in \mathbb{Z}$ for infinite ordered groups.

Proof. It is trivial to see that ϕ is bijective. Also $\forall a^j, a^k \in G$:

$$\phi(a^j \circ a^k) = \phi(a^{j+k}) = b^{j+k} = b^j * b^k = \phi(a^j) * \phi(a^k) \quad (14.4.6)$$

as required. ■

14.5 Standard groups

Definition 15.22 (*Cyclic group of order n*)

We denote the standard, abstract **cyclic group of order n** as C_n .

Example. Find an isomorphism from $(\mathbb{Z}_4, +_4)$ to (\mathbb{Z}_5, \times_5)

We know that $(\mathbb{Z}_4, +_4)$ is generated by 1 and (\mathbb{Z}_5, \times_5) by 2, so we can define:

$$\phi : (\mathbb{Z}_4, +_4) \longrightarrow (\mathbb{Z}_5, \times_5) \quad (14.5.1)$$

$$0 \mapsto 1 \quad (14.5.2)$$

$$1 \mapsto 2 \quad (14.5.3)$$

$$2 \mapsto 4 \quad (14.5.4)$$

$$3 \mapsto 3 \quad (14.5.5)$$

Definition 15.23 (*Klein four-group*)

We denote the standard, abstract group of order 4 with all elements self inverse as the **Klein four-group V** .

There are several examples where the Klein-four group comes to use. For example, let us consider the Cayley tables of $(S(\square), \circ)$ and (U_8, \times_8) :

$$\begin{array}{c|cccc}
 \circ & e & a & r & s \\
 \hline
 e & e & a & r & s \\
 a & a & e & s & r \\
 r & r & s & e & a \\
 s & s & r & a & e
 \end{array}
 \quad
 \begin{array}{c|cccc}
 \times_8 & 1 & 3 & 5 & 7 \\
 \hline
 1 & 1 & 3 & 5 & 7 \\
 3 & 3 & 1 & 7 & 5 \\
 5 & 5 & 7 & 1 & 3 \\
 7 & 7 & 5 & 3 & 1
 \end{array}$$

$(S(\square), \circ)$ (U_8, \times_8)

We can see that they both share the same structure, and therefore are isomorphic to each other.

14.6 Direct product of groups

Definition 15.24 (*Direct product of two groups*)

Given two groups $(G, *_G)$ and $(H, *_H)$, then we may define their **direct product**, denoted as $(G \times H, *)$, where:

$$G \times H = \{(g, h) : g \in G, h \in H\} \quad (14.6.1)$$

is the **cartesian product** of G, H , equipped with the operation

$$(g_1, h_1) * (g_2, h_2) = (g_1 *_G g_2, h_1 *_H h_2) \quad (14.6.2)$$

for $g_1, g_2 \in G$ and $h_1, h_2 \in H$.

We can easily see that this new group $G \times H$ does indeed satisfy the group axioms. $*$ is certainly binary. Indeed, for $(g_1, h_1), (g_2, h_2) \in G \times H$:

$$(g_1, h_1) * (g_2, h_2) = (g_1 *_G g_2, h_1 *_H h_2) \in G \times H \quad (14.6.3)$$

since G and H are closed under $*_G$ and $*_H$ respectively.

Associativity can be proven similarly.

Also, the identity element is (e_G, e_H) , since $\forall (g, h) \in G \times H$:

$$(e_G, e_H) * (g, h) = (e_G *_G g, e_H *_H h) = (g, h) \quad (14.6.4)$$

and

$$(g, h) * (e_G, e_H) = (g *_G e_G, h *_H e_H) = (g, h) \quad (14.6.5)$$

as desired.

Finally, the inverse of (g, h) is (g^{-1}, h^{-1}) , since:

$$(g, h) * (g^{-1}, h^{-1}) = (g *_G g^{-1}, h *_H h^{-1}) = (e_G, e_H) \quad (14.6.6)$$

and

$$(g^{-1}, h^{-1}) * (g, h) = (g^{-1} *_G g, h^{-1} *_H h) = (e_G, e_H) \quad (14.6.7)$$

as desired. Taking the direct product of two groups can prove to be very useful. For example, suppose we have two independent figures \mathcal{F} and \mathcal{F}' . Then, the symmetries of this overall system form the group $S(\mathcal{F}) \times S(\mathcal{F}')$.

Proposition 15.25 (Isomorphic group products)

$$C_n \times C_m \cong C_{nm} \text{ iff } \text{GCD}(m, n) = 1.$$

Proof. Suppose that $\text{GCD}(m, n) = 1$, and let $C_n = \langle a \rangle$, $C_m = \langle b \rangle$, and $\text{ord}((a, b)) = k$ so that:

$$(a, b)^k = (a^k, b^k) = e \quad (14.6.8)$$

which can only be the case if k is a common multiple of $m = \text{ord}(a)$ and $n = \text{ord}(b)$. Since k must be the minimum such integer, we require $\text{LCM}(m, n) = k$. Using the well known relation that:

$$\text{LCM}(m, n) = \frac{n \cdot m}{\text{GCD}(m, n)} = n \cdot m = k \quad (14.6.9)$$

hence the order of (a, b) is the product of the order of a and b .

Recall that $\langle (a, b) \rangle \leq C_n \times C_m$, so that $|\langle (a, b) \rangle| = \text{ord}((a, b)) = n \cdot m$. However, $|C_n \times C_m| = n \cdot m$ as well, implying that $\langle (a, b) \rangle = C_n \times C_m$. It is also immediate that $\langle (a, b) \rangle \cong C_{nm}$, so that $C_n \times C_m \cong C_{nm}$ as desired.

Now suppose that $\text{GCD}(m, n) \neq 1$, so that $k \neq n \cdot m$. Hence, C_{nm} is of order $n \cdot m$, whereas $C_n \times C_m$ is of order k . Two groups cannot be isomorphic if they have different orders, and consequently $C_{nm} \not\cong C_n \times C_m$.

■

Theorem 15.26 (Direct product theorem)

Let $H, F \leq (G, *)$, and suppose $\forall h \in H, f \in F, g \in G$:

- (i) $H \cap F = \{e\}$
- (ii) $hf = fh$
- (iii) $g = hf$

Then $G \cong H \times F$.

Proof. We will prove that:

$$\phi : H \times F \rightarrow G \quad (14.6.10)$$

$$(h, g) \mapsto h * g \quad (14.6.11)$$

is an isomorphism. Indeed,

$$\phi((h_1, f_1) * (h_2, f_2)) = \phi(h_1 * h_2, f_1 * f_2) \quad (14.6.12)$$

$$= h_1 * h_2 * f_1 * f_2 \quad (14.6.13)$$

$$= (h_1 * f_1) * (h_2 * f_2) \quad (14.6.14)$$

$$= \phi(h_1, f_1) * \phi(h_2, f_2) \quad (14.6.15)$$

where we used the commutativity of h, f in the third line.

Also, ϕ is injective. Indeed:

$$\phi(h_1, f_1) = \phi(h_2, f_2) \implies h_1 * f_1 = h_2 * f_2 \implies h_1 * h_2^{-1} = f_2 * f_1^{-1} \quad (14.6.16)$$

Now by closure, the LHS belongs to H , and the RHS belongs to F , hence $h_1 * h_2^{-1} \in H$ and $h_1 * h_2^{-1} \in F$, consequently $h_1 * h_2^{-1} \in H \cap F \implies h_1 * h_2^{-1} = e \implies h_2 = h_1$. Similar argument gives $f_2 = f_1$. Hence, $(h_1, f_1) = (h_2, f_2)$ as desired.

To prove surjectivity, note that if $g \in G$, then $\exists h, f$ such that $g = hf = \phi(h, f)$ by assumption.

In conclusion, ϕ is a bijective homomorphism, hence an isomorphism. ■

Unit B3: Permutations

15.1 Permutations

Definition 16.1 (*Permutation*)

A **permutation** of a finite set S is a bijective map $\sigma : S \rightarrow S$. The set of all permutations of this set is denoted $\text{Sym}(S)$.

We use the typical two-line notation to denote a permutation:

$$\sigma \leftrightarrow \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 2 & 3 \end{pmatrix} \quad (15.1.1)$$

A more convenient and effective notation is the **cycle form**. Indeed starting from 1 and looking at where each element gets mapped we find:

$$1 \xrightarrow{\sigma} 4 \xrightarrow{\sigma} 3 \xrightarrow{\sigma} 2 \xrightarrow{\sigma} 1 \quad (15.1.2)$$

or more concisely as:

$$\sigma = (1 \ 4 \ 3 \ 2) \quad (15.1.3)$$

However it is not always possible to write a permutation in one single cycle. Indeed some permutations map only some symbols in each cycle:

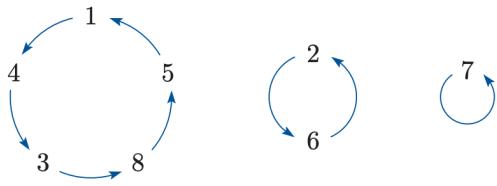
$$\sigma_g = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 6 & 8 & 3 & 1 & 2 & 7 & 5 \end{pmatrix} \quad (15.1.4)$$

then we have three **disjoint** cycles (disjoint because every member appears only in one cycle):

We can then write:

$$\sigma_g = (1 \ 4 \ 3 \ 8 \ 5)(2 \ 6)(7) \quad (15.1.5)$$

and we say that it is a product of three cycles.

**Figure 15.1.** Cycles of σ_g **Definition 16.2 (Permutation cycles)**

A permutation σ of a set S is said to be cyclic if there exist $a_i \dots a_k \in \{1, 2, \dots, n\}$ such that:

$$a_i\sigma = a_{i+1} \quad \text{for } 1 \leq i < k \quad (15.1.6)$$

$$a_k\sigma = a_1 \quad (15.1.7)$$

and is denoted in **cycle form** as $(a_1 \ \sigma a_1 \ \sigma^2 a_1 \ \dots \ \sigma^{k-1} a_1)$. We refer to this cycle as a **k -cycle** because it has order k . Two cycles are disjoint if they do not share any common elements.

The first line, $a_i\sigma = a_{i+1}$ for $1 \leq i < k$, tells us that if some a_1 gets mapped to a_2 by σ , then a_2 itself will get mapped to a_3 and so forth until we reach a_k . Here, the cycle restarts, so we must have that a_k gets mapped to a_1 , which is encapsulated in the second line.

For example, $\sigma = (1 \ 4 \ 3 \ 8 \ 5)$ is a cycle. Indeed, defining $a_1 = 1$ then:

$$a_1\sigma = 4 = a_2, \quad (15.1.8)$$

$$a_2\sigma = 3 = a_3, \quad (15.1.9)$$

$$a_3\sigma = 8 = a_4, \quad (15.1.10)$$

$$a_4\sigma = 5 = a_5 \quad (15.1.11)$$

$$a_5\sigma = 1 = a_1 \quad (15.1.12)$$

so here $k = 5$, thus we have a 5-cycle.

Proposition 16.3 (Commutativity product of disjoint cycles)

The product of disjoint cycles is **commutative**.

Proof. Let $\sigma_a = (a_1 \dots a_k)$ and $\sigma_b = (b_1 \dots b_l)$ so that $a_i \neq b_i$. Then applying $\sigma_b \sigma_a$:

$$a_i \sigma_a \sigma_b = a_{i+1} \sigma_b = a_{i+1} \quad \text{for } i < k \quad (15.1.13)$$

$$a_k \sigma_a \sigma_b = a_1 \sigma_b = a_1 \quad (15.1.14)$$

$$b_i \sigma_a \sigma_b = b_i \sigma_b = b_{i+1} \quad \text{for } i < l \quad (15.1.15)$$

$$b_l \sigma_a \sigma_b = b_l \sigma_b = b_1 \quad (15.1.16)$$

Similarly applying $\sigma_b \sigma_a$ one finds:

$$a_i \sigma_b \sigma_a = a_i \sigma_a = a_{i+1} \quad \text{for } i < k \quad (15.1.17)$$

$$a_k \sigma_b \sigma_a = a_k \sigma_a = a_1 \quad (15.1.18)$$

$$b_i \sigma_b \sigma_a = b_{i+1} \sigma_a = b_{i+1} \quad \text{for } i < l \quad (15.1.19)$$

$$b_l \sigma_b \sigma_a = b_1 \sigma_a = b_1 \quad (15.1.20)$$

■

An immediate consequence of Proposition 16.3 is that, if a permutation σ can be expressed as a product of disjoint cycles σ_i :

$$\sigma = \prod_{i=1}^n \sigma_i \implies \sigma^k = \prod_{i=1}^n \sigma_i^k \quad (15.1.21)$$

It turns out that this process of writing permutations as products of disjoint cycles can be done for any cycle, and this process is well-defined. In other words, the cycle form is uniquely determined for any permutation.

Theorem 16.4 (*Existence and uniqueness cycle form*)

Every permutation can be written in a unique cycle form (aside the choice of starting symbol and order in which the symbols are listed).

Proof. Proof of existence

Consider a permutaton $\sigma \in \text{Sym}(\{1, 2, \dots, n\})$ and the infinite sequence:

$$a_1, a_1 \sigma, a_1 \sigma^2, a_1 \sigma^3 \dots \quad (15.1.22)$$

where $a_1 \in \{1, 2, \dots, n\}$. Because $\{1, 2, \dots, n\}$ has finite cardinality, and the sequence continues infinitely, we must have some repetition $a_1 \sigma^i = a_1 \sigma^j$ for some $i < j$ which implies $a_1 \sigma^{j-i} = a_1$. If we let $k_1 = j - i$ be the smallest integer such that $a_1 \sigma^{k_1} = a_1$ then we denote $\{a_1, a_1 \sigma, \dots, a_1 \sigma^{k_1-1}\}$ the **orbit** of a_1 which is our first cycle.

Now if $k_1 = n$ then the permutation σ is a cycle, and we are done. Otherwise, we choose a_1 not in the orbit of a_1 and write its orbit.

The orbits of a_1 and a_2 are disjoint since σ is bijective. Indeed, if $\sigma^i a_1 = \sigma^j a_2$ then $a_1 = \sigma^{j-i} a_2$, implying that a_1 belongs to the orbit of a_2 , a contradiction.

Repeating this process, since the set S is finite eventually we exhaust the number of symbols and find:

$$\sigma = (a_1 \ a_1\sigma \dots \ a_1\sigma^{k_1-1})(a_2 \ a_2\sigma \dots \ a_2\sigma^{k_2-1}) \dots (a_r \ a_R\sigma \dots \ a_r\sigma^{k_r-1}) \quad (15.1.23)$$

where r is the different number of orbits.

Proof of uniqueness Suppose now that we can write the permutation as two distinct products of disjoint cycles:

$$\sigma = \rho_1 \rho_2 \dots \rho_k = \tau_1 \tau_2 \dots \tau_l \quad (15.1.24)$$

Then $a_1 \in \{1, 2, \dots, n\}$ appears exactly once in ρ_i and τ_j and as they are disjoint we reorder (through commutativity) the order of cycles so that $i = j = 1$. Hence WLOG assume that 1 appears in τ_1 and ρ_1 .

We can then show that:

$$\rho_1 = (1 \ 1\sigma \ 1\sigma^2 \ \dots \ 1\sigma^{k-1}) = \tau_1 \quad (15.1.25)$$

an repeating this for all other ρ and τ we get the desired result. ■

During this proof we encountered the important concept of an orbit which we shall revisit more in depth later when studying group actions:

Definition 16.5 (*Orbit of a permutation element*)

The orbit of some element $a_1 \in S$ in a permutation $\sigma \in \text{Sym}(S)$ is the set $\{a_1, \sigma a_1 \dots \sigma^{k_1-1} a_1\}$ where k_1 is the smallest integer such that $\sigma^{k_1} a_1 = a_1$.

Strategy. (*Composing permutations*)

To find $\sigma_g \circ \sigma_f$ in cycle form:

1. Start with 1 and find the symbol of 1 under σ_f , and then find the image of that symbol under σ_g , and denote it as $x = \sigma_g \sigma_f 1$ so that $\sigma_g \circ \sigma_f = (1 \ x \dots)$.
2. Start with the symbol x and repeat the process.
3. Continue repeating the process until you reach the original symbol 1. The cycle is then complete.
4. Choose the smallest symbol not placed in the cycle, and repeat steps 1-3 again until the second cycle is complete.
5. Continue until every symbol has been placed.

Example. Write in cycle form $(1\ 4\ 6)(3\ 5) \circ (1\ 5\ 3\ 2\ 4) \circ (1\ 2)(3\ 5)(4\ 6)$.

We start with 1, which gets mapped to 2, then to 4 and finally to 6, so $1 \mapsto 6$.

Now we see that 6 gets mapped to 4, then to 1 and finally to 4, so $6 \mapsto 4$.

Now we see that 4 gets mapped to 6 and then doesn't get mapped and finally to 1, so $4 \mapsto 1$.

This completes the first cycle $(1\ 6\ 4)$.

We start with 2, which gets mapped to 1, then to 5 and finally to 3, so $2 \mapsto 3$.

Now we see that 3 gets mapped to 5, then to 3 and finally to 5 so $3 \mapsto 5$.

Now we see that 5 gets mapped to 3, then to 2 and then doesn't get mapped, so $5 \mapsto 2$.

This completes the second cycle $(2\ 3\ 5)$.

So we may write:

$$(1\ 4\ 6)(3\ 5) \circ (1\ 5\ 3\ 2\ 4) \circ (1\ 2)(3\ 5)(4\ 6) = (1\ 6\ 4)(2\ 3\ 5) \quad (15.1.26)$$



15.2 Permutation groups

Theorem 16.6 (*Symmetric group*)

The set S_n of all permutations of $\{1, 2, 3, \dots, n\}$ forms a group under \circ called the **symmetric group of degree n**

Proof. We check that the group axioms hold:

(Closure) Consider $\sigma_g \circ \sigma_f \in S_n$. Since σ_g and σ_f are bijective maps mapping $\{1, 2, \dots, n\}$ to itself, then $\sigma_g \circ \sigma_f$ and $\sigma_f \circ \sigma_g$ must necessarily also map S_n to itself and be bijective. So, $\sigma_g \circ \sigma_f \in S_n$ as required.

(Associativity) Composition is associative.

(Identity) The identity permutation e is an identity, since its action is to map every symbol of $\{1, 2, \dots, n\}$ to itself.

(Inverse) Since σ_f is bijective, it must have an inverse σ_f^{-1} which is also a permutation since it too maps $\{1, 2, \dots, n\}$ to itself.



It is important to notice the difference between the order and degree of S_n . The order is, as always, the number of permutations S_n contains whereas its degree is how many symbols each of its permutations permute.

Theorem 16.7 (Properties of S_n)

The order $|S_n|$ of S_n is $n!$, and for $n \geq 3$, S_n is non-abelian.

Proof. Firstly, for $\sigma \in S_n$, there are n different possibilities for 1σ , and since σ is bijective, $1\sigma \neq 2\sigma$ and thus there are $n - 1$ possibilities for 2σ . We can keep doing this to find:

$$n \times (n - 1) \times (n - 2) \dots \times 2 \times 1 = n! \quad (15.2.1)$$

are all the possible permutations. Now if $x_1, x_2, x_3 \in \{1, 2, \dots, n\}$ are distinct then we can define:

$$\sigma_1 : x_1 \mapsto x_1, x_2 \mapsto x_3, x_3 \mapsto x_2 \quad (15.2.2)$$

$$\sigma_2 : x_1 \mapsto x_2, x_2 \mapsto x_1, x_3 \mapsto x_3 \quad (15.2.3)$$

We then find that:

$$\sigma_2 \circ \sigma_1 : x_1 \mapsto x_2, x_2 \mapsto x_3, x_3 \mapsto x_1 \quad (15.2.4)$$

$$\sigma_1 \circ \sigma_2 : x_1 \mapsto x_3, x_2 \mapsto x_1, x_3 \mapsto x_2 \quad (15.2.5)$$

which are not the same. Thus for sets S of cardinality greater than or equal to 3 the corresponding $\text{Sym}(S)$ is non-abelian. ■

Definition 16.8 (Same cycle structure)

Two permutations in S_n have the same cycle structure if their cycle forms contain the same number of disjoint k -cycles for each k .

So for example $(1\ 2\ 4)(3\ 8)(4\ 5)$ and $(1\ 7)(2\ 8\ 3)(4\ 5)$ have the same cycle structure since they each consist of a 3-cycle, two **transposition** (2-cycle) and a 1-cycle (which is not shown in cycle form).

Proposition 16.9 (Order of a permutation)

The order of a permutation $\sigma = \rho_1 \rho_2 \dots \rho_k$ where ρ_i are disjoint cycles of order l_i is:

$$\text{ord}(\sigma) = \text{LCM}(l_1, l_2, \dots, l_k) \quad (15.2.6)$$

so that the order of a k -cycle is k .

Proof. Let $n = \text{ord}(\sigma)$. Then, since ρ_i are disjoint, we can use 16.1.21 to write:

$$\sigma^n = \rho_1^n \rho_2^n \rho_3^n \dots \rho_k^n = e \quad (15.2.7)$$

Now since ρ_i are disjoint, they permute different sets, and consequently their product can only be equal to e if each of them are equal to e :

$$\rho_i^n = e, \forall i \quad (15.2.8)$$

so that $n|l_i$. Since n must be the smallest such integer, it follows that $n = \text{LCM}(l_1, l_2, \dots, l_k)$ as desired. \blacksquare

We can use cycle form to denote symmetries of figures as well. For example the symmetry:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \end{pmatrix} \quad (15.2.9)$$

can be written as $(2\ 4)$ in cycle form, which is much more concise.

Example. Write all the symmetries of an octahedron.

The octahedron has six direct symmetries: we can rotate it about the vertical line through the vertices 4,5 through $0, 2\pi/3, 4\pi/3$, or turn the octahedron upside down and repeat the same process. The octahedron also has at least one indirect symmetry, name a reflection in the plane through vertices 1,2,3. Hence there are 6 indirect symmetries, and 12 symmetries in total.

We first start by writing all direct symmetries of the equilateral triangle with vertices 1,2,3:

$$e, (1\ 2\ 3), (1\ 3\ 2), (1\ 2), (1\ 3), (2\ 3) \quad (15.2.10)$$

We can then compose with the reflection $(4\ 5)$ to find:

$$(4\ 5), (1\ 2\ 3)(4\ 5), (1\ 3\ 2)(4\ 5), (1\ 3)(4\ 5), (2\ 3)(4\ 5) \quad (15.2.11)$$

These are twelve distinct symmetries, and therefore we have found all the symmetries. \blacktriangleleft

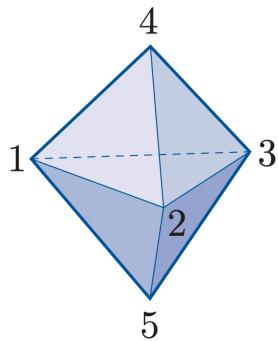
15.3 Even and Odd symmetries

Strategy. (*Expressing cycles as composite of transpositions*)

Consider a cycle $(a_1\ a_2\ a_3\dots a_r)$, then we can express:

$$(a_1\ a_2\ a_3\dots a_r) = (a_1\ a_r) \circ (a_1\ a_{r-1}) \circ \dots \circ (a_1\ a_2) \quad (15.3.1)$$

Proof. Firstly, consider a_1 . It gets mapped to a_2 . a_2 then gets mapped to itself, since it does not appear in any other transposition (they are disjoint). So overall a_1 gets mapped to a_2 .

**Figure 15.2.** Octahedron

Now consider the a_s where $2 \leq s \leq r - 1$. We then see that $(a_1 a_2), \dots, (a_1 a_{s-1})$ all map a_s to itself. The next transposition $(a_1 a_s)$ maps a_s to a_1 . Then the next transposition $(a_1 a_{s+1})$ maps a_1 to a_{s+1} . Finally, all the successive transpositions $(a_1 a_{s+2}), \dots, (a_1 a_r)$ map a_{s+1} to itself. So we find that overall a_s gets mapped to a_{s+1} .

Next we consider a_r . By the same argument as before all transpositions $(a_1 a_2), \dots, (a_1 a_{r-1})$ map a_r to itself. The final transpositions $(a_1 a_r)$ map a_r to a_1 as required.

So we find that $a_s \mapsto a_{s+1}$ for all $1 \leq s < r$ and $a_r \mapsto a_1$, which defines the cycle $(a_1 a_2 a_3 \dots a_r)$ as in definition 16.2. ■

Example. We can express $(2 \ 4 \ 3 \ 5) = (2 \ 5) \circ (2 \ 3) \circ (2 \ 4)$. ◀

Notice that there are several ways we can express a permutation as a composite of transpositions. Indeed in the previous example we could have written:

$$(2 \ 4 \ 3 \ 5) = (4 \ 3 \ 5 \ 2) = (4 \ 2) \circ (4 \ 5) \circ (4 \ 3) \quad (15.3.2)$$

$$= (2 \ 4) \circ (5 \ 4) \circ (3 \ 4) \quad (15.3.3)$$

Notice however that the number of transpositions is always even. Indeed it turns out that if a permutation can be expressed as a composition of an even number of transpositions, then it can only be expressed as a composite of even transpositions.

Definition 16.9 (*Parity of permutation*)

A permutation is **even** if it can be expressed as a composite of an even number of transpositions.

A permutation is **odd** if it can be expressed as a composite of an odd number of transpositions.

We have that:

Theorem 16.10 (Parity Theorem)

A permutation cannot be expressed as both a composite of an even number of transpositions and a composite of an odd number of transpositions.

Proof. Let x_1, \dots, x_n and consider the Vandermonde polynomial ¹

$$P = \prod_{1 \leq i < j \leq n} (x_j - x_i) \quad (15.3.4)$$

We now choose a permutation $\sigma \in S_n$, and define the function f_σ as:

$$f_\sigma(P) = \prod_{1 \leq i < j \leq n} (x_{\sigma(j)} - x_{\sigma(i)}) \quad (15.3.5)$$

which reshuffles the terms in the normal Vandermonde polynomial.

Lemma. Let τ be a transposition, then $f_\tau(P) = -P$.

Proof. Let's consider the action of a transposition $\tau = (a b)$ on P .

For $j, i \neq a, b$ in any order, the factor $(x_j - x_i)$ is left unchanged.

Now let us consider a pair with exactly one index equal to a or b (we assume WLOG that $a < b$).

Then if the other index j is between a and b , $(x_j - x_a) \mapsto -(x_b - x_j)$ and $(x_b - x_j) \mapsto -(x_j - x_a)$. However these two signs cancel each other out, so no net change.

If the other index i is not between a and b , then it can be smaller than a or larger than b . In the first case, $(x_a - x_i) \mapsto (x_b - x_i)$. In the former case, $(x_i - x_a) \mapsto (x_i - x_b)$. In both cases the sign does not change.

Finally, if both indices are in $\{a, b\}$ then $x_b - x_a \mapsto -(x_b - x_a)$.

So overall there is only one sign change due to when both indices correspond to a and b . So if τ is a transposition, then $f_\tau(P) = -P$ and consequently $f_\tau(-P) = P$. ■

Now take an arbitrary permutation σ , and express it as a product of transpositions τ_i and ρ_i in two different ways:

$$\sigma = \tau_1 \cdots \tau_r = \rho_1 \cdots \rho_s \quad (15.3.6)$$

¹For $n = 4$, we would have $P = (x_2 - x_1)(x_3 - x_1)(x_4 - x_1)(x_3 - x_2)(x_4 - x_2)(x_4 - x_3)$.

Then

$$f_\sigma(P) = f_{\tau_1 \dots \tau_r}(P) = (f_{\tau_1} \circ \dots \circ f_{\tau_r})(P) = (-1)^r P \quad (15.3.7)$$

$$f_\sigma(P) = f_{\rho_1 \dots \rho_s}(P) = (f_{\rho_1} \circ \dots \circ f_{\rho_s})(P) = (-1)^s P \quad (15.3.8)$$

Therefore, $(-1)^r P = (-1)^s P$, so r and s have the same parity: both odd, or both even. ■

Proposition 16.11 (Parity of k -cycles)

For $\sigma \in S_n$ where σ is a k -cycle:

$$\sigma \text{ is } \begin{cases} \text{even permutation, if } k \text{ is odd} \\ \text{odd permutation, if } k \text{ is even} \end{cases} \quad (15.3.9)$$

Proof. This follows immediately from equation 16.3.1 ■

We note immediately then that the parity of the composite of two permutations can be found by simply summing their parity.

Strategy. (Finding the parity of any permutation)

1. Express the permutation as a composite of cycles
2. find the parity of each k -cycle following

$$\begin{cases} \text{even permutation, if } k \text{ is odd} \\ \text{odd permutation, if } k \text{ is even} \end{cases} \quad (15.3.10)$$

3. Combine the parities of each cycle following the Cayley table below:

+	even	odd
even	even	odd
odd	odd	even

Proposition 16.12 (Parity of inverse)

The inverse of a permutation has the same parity as the permutation.

Proof. We see that e is an even permutation, so if $f \circ f^{-1} = e$ then we must have that f and f^{-1} have the same parity. ■

Proposition 16.13 (Alternating group of order n)

The group A_n of all even permutations of $\{1, 2, \dots, n\}$ is called the **alternating group of order n** and $A_n \leq S_n$.

Proof. We check the subgroup properties:

- (i) **Closure:** the composite of two even permutations is even, so closure is satisfied.
- (ii) **Identity:** the identity permutation e is even, and thus belongs to A_n .
- (iii) **Inverses:** by theorem 16.10, the inverse of an even permutation is even.

■

For example, let us try to find all elements of A_4 . Their structures must be:

$$e, (\underline{\quad \quad})(\underline{\quad \quad}), (\underline{\quad \quad \quad \quad}) \quad (15.3.11)$$

and we must fill the gaps with $\{1, 2, 3, 4\}$.

Note that there are only 3 cycle structures of the form $(\underline{\quad \quad})(\underline{\quad \quad})$. Indeed, WLOG place 1 in the first place, then we can place 3 elements in the second place. The other transposition is then given immediately. So there are three different ways.

Instead, there are 8 cycle structures of the form $(\underline{\quad \quad \quad \quad})$. For these, we place WLOG 1 in the first place so that there are 3 possible elements in the second place, and 2 possible in the third, the fourth is then immediate. So there are eight different ways.

In total there are then 12 different elements in A_4 , which is exactly half the order of S_4 interestingly.

This turns out not to be a coincidence. Indeed, we have the following general result:

Theorem 16.14 (Order of A_n)

The order of the alternating group is $|A_n| = \frac{1}{2}n!$ for $n \geq 2$.

Proof. Suppose S_n has r even permutations and s odd permutations.

We first prove that $r \leq s$. Let $f_1, \dots, f_r \in A_n$ then consider:

$$(1 \ 2) \circ f_1, (1 \ 2) \circ f_2, \dots, (1 \ 2) \circ f_r \quad (15.3.12)$$

these are all distinct odd permutations. So there are at least r odd permutations in S_n .

A similar argument uses $g_1 \dots g_r$ odd permutations composed with $(1 \ 2)$ to prove that $s \leq r$.

Since $s \leq r$ and $r \leq s$, we have that $s = r$. So exactly half of the permutations in S_n are in A_n , thus $|A_n| = \frac{1}{2}n!$. ■

15.4 Conjugacy of S_n

Consider the permutation x of some set S with $i, j \in S$. Then consider another permutation g which relabels S . Our question is to know what the permutation x looks like with the relabelled set S' .

Looking at the diagram below:

$$\begin{array}{ccc} i & \xrightarrow{x} & j \\ g^{-1} \uparrow & & \downarrow g \\ g(i) & \xrightarrow{y} & g(j) \end{array}$$

we clearly see that $(y \circ g)(i) = (g \circ x \circ g^{-1})(i)$ and by the cancellation rule we find that $y = g \circ x \circ g^{-1}$.

Definition 16.15 (Conjugate permutations)

The permutation σ is the conjugate of ρ in S_n if there exists a permutation τ such that:

$$\sigma = \tau \circ \rho \circ \tau^{-1} \quad (15.4.1)$$

We then say that τ is a **conjugating permutation** of ρ to σ , and that σ is the **conjugate** of ρ by τ .

Strategy. (Finding a conjugating permutation)

1. Align the cycles of σ and ρ so that cycles of same order correspond:

$$\begin{aligned} x &= (* * \dots *) (* * \dots *) \dots (*)(*) \Big) \tau \\ y &= (* * \dots *) (* * \dots *) \dots (*)(*) \Big) \tau \end{aligned} \quad (15.4.2)$$

2. Read off the two line form of the permutation τ .

Example. Let $\sigma = (1\ 2\ 4)(3\ 5)$ and $\rho = (1\ 4)(2\ 5\ 3)$ in S_5 . Find three permutations $g \in S_5$ conjugating σ to ρ .

We can write that:

$$\begin{aligned} x &= (1\ 2\ 4)(3\ 5) \Big) \tau \\ y &= (2\ 5\ 3)(1\ 4) \Big) \tau \end{aligned} \quad (15.4.3)$$

and read off the conjugating permutation $\tau = (1\ 2\ 5\ 4\ 3)$.

Alternatively, we can rewrite $(2\ 5\ 3)$ as $(3\ 2\ 5)$ and find:

$$\begin{aligned} x &= (1\ 2\ 4)(3\ 5) \\ y &= (3\ 2\ 5)(1\ 4) \end{aligned} \quad \swarrow \tau \tag{15.4.4}$$

whose corresponding conjugate permutation is $\tau = (1\ 3)(4\ 5)$.

Finally we can rewrite $(2\ 5\ 3)$ as $(3\ 5\ 2)$ and find:

$$\begin{aligned} x &= (1\ 2\ 4)(3\ 5) \\ y &= (3\ 5\ 2)(1\ 4) \end{aligned} \quad \swarrow \tau \tag{15.4.5}$$

whose corresponding conjugate permutation is $\tau = (1\ 3)(2\ 5\ 4)$ ◀

Now consider the action of a conjugating permutation not on a single permutation, but on every permutation in a subgroup.

Let $H \leq S_n$ and let $g \in S_b$, then we will denote:

$$g \circ H \circ g^{-1} = \{g \circ h \circ g^{-1} : h \in H\} \tag{15.4.6}$$

So it suffices to substitute every element in H using g .

If we let for example $H = \langle(1\ 2\ 4\ 5)\rangle$ then:

$$H = \{e, (1\ 2\ 4\ 5), (1\ 2\ 4\ 5)^2, (1\ 2\ 4\ 5)^3\} \tag{15.4.7}$$

$$= \{e, (1\ 2\ 4\ 5), (1\ 4)(2\ 5), (1\ 5\ 4\ 2)\} \tag{15.4.8}$$

Then, if we let $g = (3\ 5)$ we find that:

$$g \circ H \circ g^{-1} = \{e, (1\ 2\ 4\ 3), (1\ 4)(2\ 3), (1\ 3\ 4\ 2)\} \tag{15.4.9}$$

Theorem 16.16 (Conjugate subgroups)

Let $H \leq S_n$ and let $g \in S_n$. Then $g \circ H \circ g^{-1}$ is also a subgroup of S_n .

Proof.

Closure: consider any two elements $h, k \in g \circ H \circ g^{-1}$. Then, $\exists h', k'$ such that $h = g \circ h' \circ g^{-1}$ and $k = g \circ k' \circ g^{-1}$. Thus:

$$h \circ k = (g \circ h' \circ g^{-1}) \circ (g \circ k' \circ g^{-1}) \tag{15.4.10}$$

$$= g \circ h' \circ g^{-1} \circ g \circ k' \circ g^{-1} \tag{15.4.11}$$

$$= g \circ (h' \circ k) \circ g^{-1} \tag{15.4.12}$$

$$= g \circ l \circ g^{-1} \tag{15.4.13}$$

where $l = h' \circ k \in H$ since subgroups are closed. Hence $h \circ k \in g \circ H \circ g^{-1}$ as required.

Identity: let the identity element in H be e_H . Then:

$$e_H = g \circ g^{-1} = g \circ e_H \circ g^{-1} \in g \circ H \circ g^{-1} \quad (15.4.14)$$

as required.

Inverses Let $h \circ h^{-1} = e_H$ for all $h \in H$. Then, the inverse of $g \circ h \circ g^{-1}$ is $g \circ h^{-1} \circ g^{-1}$. Indeed:

$$g \circ h \circ g^{-1} \circ g \circ h^{-1} \circ g^{-1} = g \circ h \circ h^{-1} \circ g^{-1} = g \circ g^{-1} = e_H \quad (15.4.15)$$

Let us now check that the inverse belongs to $g \circ H \circ g^{-1}$. This is clearly true, since $h^{-1} \in H$ due to the inverse property of subgroups.

■

15.5 Subgroups of S_4

We will now tackle the problem of finding all subgroups of S_4 . To do so we will find cyclic and all non-cyclic subgroups of S_4 .

Cyclic subgroups

The cyclic elements of S_4 must have the structures shown in Figure 16.3. We see that each

Cycle structure	Order	Elements of S_4	Description
e	1	e	identity
$(--)$	2	$(1\ 2), (1\ 3), (1\ 4), (2\ 3), (2\ 4), (3\ 4)$	transpositions
$(---)$	3	$(1\ 2\ 3), (1\ 3\ 2), (1\ 2\ 4), (1\ 4\ 2), (1\ 3\ 4), (1\ 4\ 3), (2\ 3\ 4), (2\ 4\ 3)$	3-cycles
$(----)$	4	$(1\ 2\ 3\ 4), (1\ 2\ 4\ 3), (1\ 3\ 2\ 4), (1\ 3\ 4\ 2), (1\ 4\ 2\ 3), (1\ 4\ 3\ 2)$	4-cycles
$(--)(--)$	2	$(1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3)$	products of 2-cycles

Figure 15.3. Cycle structures of elements in S_4

element in S_4 has order 1,2,3,4 so the order of each subgroup must also be 1,2,3 or 4.

The cyclic subgroup of order 1 is obviously $\{e\}$.

The cyclic subgroups of order 2 are the identity permutation with one permutation of order 2:

$$\{e, (1 2)\}, \{e, (1 3)\}, \{e, (1 4)\}, \{e, (2 3)\}, \{e, (2 4)\}, \{e, (3 4)\} \quad (15.5.1)$$

and:

$$\{e, (1 2)(3 4)\}, \{e, (1 3)(2 4)\}, \{e, (1 4)(2 3)\} \quad (15.5.2)$$

which are 9 in total.

To find all cyclic subgroups of order 3, note that $(1 2 3)$ and $(1 3 2)$ all generate the same subgroup $\{e, (1 2 3), (1 3 2)\}$. Similarly, the other six 3-cycles each couple up in a similar way. So the cyclic groups are:

$$\langle\langle(1 2 3)\rangle, \langle(1 3 4)\rangle, \langle(1 4 2)\rangle, \langle(2 3 4)\rangle\rangle \quad (15.5.3)$$

The cyclic subgroups of order 4 can be found similarly. $(1 2 3 4)$ generates the following set:

$$\langle(1 2 3 4)\rangle = \{e, (1 2 3 4), (1 3)(2 4), (1 4 3 2)\} \quad (15.5.4)$$

We now choose a permutation of order 4 that is not in this list, and finds its generator:

$$\langle(1 2 4 3)\rangle = \{e, (1 2 4 3), (1 4)(2 3), (1 3 4 2)\} \quad (15.5.5)$$

Repeat this process one final time:

$$\langle(1 3 2 4)\rangle = \{e, (1 3 2 4), (1 2)(3 4), (1 4 2 3)\} \quad (15.5.6)$$

We see that all six permutations of order 4 have been found, so we found all the cyclic subgroups of S_4 of order 4.

So in conclusion:

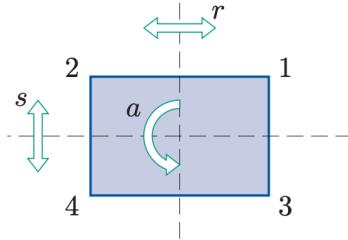
Order	Number of cyclic subgroups
1	1
2	9
3	4
4	3

with 16 total cyclic subgroups.

Non-cyclic subgroups

We now try to find non-cyclic subgroups of S_4 . We can do so by drawing a figure labels 1, 2...n. We can then find the symmetry group of the figure, which is a subgroup of S_n .

For example, if we draw and label the rectangle as shown below:



Then the symmetry group is:

$$\{e, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3)\} \quad (15.5.7)$$

which is not cyclic. Indeed it has 4 elements, but its elements are of order 2.

We can now find other non-cyclic subgroup from the old subgroup by relabelling it through conjugating permutations.

15.6 Cayley's Theorem

We saw that the symmetry groups of most figures can be represented as permutation groups, and are therefore isomorphic.

It turns out this is true for any finite group.

Theorem 16.17 (Cayley's theorem)

Let $(G, *)$ be a finite group. For each $x \in G$, let p_x be the permutation whose two-line symbol has as its first line the column heading of the Cayley table of G , and as its second line the row labelled x in the group table.

If we let $P = \{p_x : x \in G\}$ then (P, \circ) is a permutation group isomorphic to $(G, *)$.

Proof. Let $G = \{g_1, \dots, g_n\}$ so that for each element g_i the table shows:

*	g_1	g_2	g_3	\dots	g_n
g_1					
\vdots					
x	$x * g_1$	$x * g_2$	$x * g_3$	\dots	$x * g_n$
\vdots					
g_n					

so that:

$$p_x = \begin{pmatrix} g_1 & g_2 & \dots & g_n \\ x * g_1 & x * g_2 & \dots & x * g_n \end{pmatrix} \quad (15.6.1)$$

which is a permutation since every element of G is repeated only once in the column headings and in the row labelled x .

One can then easily verify using associativity that:

$$p_x \circ p_y = \begin{pmatrix} g_1 & g_2 & \cdots & g_n \\ x * y * g_1 & x * y * g_2 & \cdots & x * y * g_n \end{pmatrix} = p_{x * y} \quad (15.6.2)$$

so that:

*	...	y	...					
:		:		:		:		
x	...	$x * y$...	p_x	...	$p_{x * y}$...	
:		:		:		:		

$(G, *)$ (P, \circ)

The two Cayley tables therefore are identical in structure, and thus (P, \circ) is a group. Not only that, it is a group isomorphic to $(G, *)$, since the map:

$$p : G \rightarrow P \quad (15.6.3)$$

$$x \mapsto p_x \quad (15.6.4)$$

Indeed, we have already verified that $p(x * y) = p(x) \circ p(y)$. Also, p is injective, since $p(x) = p(y) \implies x * g_i = y * g_i$ for all $g_i \in G$, and by the right cancellation law $x = y$. Surjectivity is trivial. ■

For example, consider the group $(\mathbb{Z}_6, +_6)$. For each $x \in \mathbb{Z}_6$, we associate a permutation p_x

$+_6$	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

Figure 15.4. Cayley table for $(\mathbb{Z}_6, +_6)$

whose two line form has its first line as the column heading and its second line as the row labelled x . So:

$$p_2 = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 4 & 5 & 0 & 1 \end{pmatrix} = (0 \ 2 \ 4)(1 \ 3 \ 5) \quad (15.6.5)$$

This gives us the permutations $p_0, p_1 \dots p_6$ for each element of \mathbb{Z}_6 , obtaining:

$$p_0 = e \quad (15.6.6)$$

$$p_1 = (0 \ 1 \ 2 \ 3 \ 4 \ 5) \quad (15.6.7)$$

$$p_2 = (0 \ 2 \ 4)(1 \ 3 \ 5) \quad (15.6.8)$$

$$p_3 = (0 \ 3)(1 \ 4)(2 \ 5) \quad (15.6.9)$$

$$p_4 = (0 \ 4 \ 2)(1 \ 5 \ 3) \quad (15.6.10)$$

$$p_5 = (0 \ 5 \ 4 \ 3 \ 2 \ 1) \quad (15.6.11)$$

Let $P = \{p_0, p_1, p_2 \dots p_5\}$. If we draw the Cayley table for (P, \circ) we find:

\circ	p_0	p_1	p_2	p_3	p_4	p_5
p_0	p_0	p_1	p_2	p_3	p_4	p_5
p_1	p_1	p_2	p_3	p_4	p_5	p_0
p_2	p_2	p_3	p_4	p_5	p_0	p_1
p_3	p_3	p_4	p_5	p_0	p_1	p_2
p_4	p_4	p_5	p_0	p_1	p_2	p_3
p_5	p_5	p_0	p_1	p_2	p_3	p_4

which is structurally identical to the table for $(\mathbb{Z}_6, +_6)$. So we can conclude that the map:

$$\phi : \mathbb{Z}_6 \rightarrow P \quad (15.6.12)$$

$$x \mapsto p_x \quad (15.6.13)$$

is an isomorphism.

Unit B4: Lagrange's Theorem and small groups

16.1 Lagrange's Theorem

Theorem 17.1 (Lagrange's Theorem)

Let G be a finite group, and let $H \leq G$. Then $\text{ord}(H)|\text{ord}(G)$, that is, the order of H divides the order of G .

Proof. Let (G, \circ) be a finite group and let $H \leq G$ so that $\text{ord}(G) = s$ and $\text{ord}(H) = r$.

We begin by writing down all the elements of H :

$$(h_1 \ h_2 \ \dots \ h_r) \quad (16.1.1)$$

Next we choose any element of G that is not included in the above array, such as g_2 , and compose it to the left with the first row.

$$\begin{pmatrix} h_1 & h_2 & \dots & h_r \\ g_2 \circ h_1 & g_2 \circ h_2 & \dots & g_2 \circ h_r \end{pmatrix} \quad (16.1.2)$$

If there are no other elements of G excluded from the array, then we are done. Otherwise, choose another element, say g_3 that is not included and compose it to the left with the first row to find:

$$\begin{pmatrix} h_1 & h_2 & \dots & h_r \\ g_2 \circ h_1 & g_2 \circ h_2 & \dots & g_2 \circ h_r \\ g_3 \circ h_1 & g_3 \circ h_2 & \dots & g_3 \circ h_r \end{pmatrix} \quad (16.1.3)$$

We repeat this process until all the elements of G have been exhausted. This must happen since G has finite order and each row introduces a new element $g_i \in G$.

At the end, we reach the following array:

$$\begin{pmatrix} h_1 & h_2 & \dots & h_r \\ g_2 \circ h_1 & g_2 \circ h_2 & \dots & g_2 \circ h_r \\ g_3 \circ h_1 & g_3 \circ h_2 & \dots & g_3 \circ h_r \\ \vdots & \vdots & & \vdots \\ g_k \circ h_1 & g_k \circ h_2 & \dots & g_k \circ h_r \end{pmatrix} \quad (16.1.4)$$

Next, we have to show that all the elements in the array are distinct. We start by showing that all the elements in a row are distinct. This is clearly true for the first row, since they are all distinct elements of H . For the k th row, we have that if for some $h_i, h_j \in H$ distinct:

$$g_k \circ h_i = g_k \circ h_j \quad (16.1.5)$$

then by the left cancellation law $h_i = h_j$ which is a contradiction.

Secondly, we show that elements in a row are not repeated in any other row. Again, we go by contradiction, and suppose that in some row l , the element $g_l \circ h_i$ is repeated as $g_k \circ h_j$ in another row k :

$$g_l \circ h_i = g_k \circ h_j \implies g_l = g_k \circ h_j \circ h_i^{-1} \quad (16.1.6)$$

By the closure property of H , $h_j \circ h_i^{-1} \in H$, which would imply that $g_l = g_k \circ h_m$ for some m and that therefore g_l belongs to the k th row. This is a contradiction, since we assumed that the rows l and k are different.

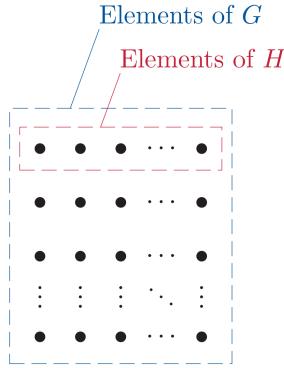


Figure 16.1. Visualization of Lagrange's proof

Thus none of the elements in each row are repeated in other rows. We can therefore conclude that the order of G is the size of the complete matrix, that is, $\text{ord}(G) = k \cdot r = k \cdot \text{ord}(H)$. It follows immediately that $\text{ord}(H) | \text{ord}(G)$. ■

Corollary.

- (i) let $g \in G$, then the $\text{ord}(g) | \text{ord}(G)$.

- (ii) let G be a group of prime order. Then G is cyclic, with every non-identity element being a generator.
- (iii) let G be a group of prime order p , then $(G, \circ) \cong (\mathbb{Z}_p, +_p)$.

Proof.

- (i) we have that $\langle g \rangle \leq G$ and $\text{ord}(g) = \text{ord}(\langle g \rangle)$ so that $\text{ord}(g) | \text{ord}(G)$.
- (ii) if G has prime order p , then for every element $g \in G$, $\langle g \rangle$ can have order 1 or p . However, only $\langle e \rangle$ has order 1, therefore $\langle x \rangle$ must have order p , and therefore generate G .
- (iii) we have that (G, \circ) is a cyclic group of order p , and that $(\mathbb{Z}_p, +_p)$ too is a cyclic group of order p . From proposition 15.21, the two groups must therefore be isomorphic.

■

16.2 Groups of small order

The goal of this section will be to justify the following classification of isomorphism classes for small groups:

Order	Standard group(s)	Properties	Further examples
1	C_1	cyclic	$(\{0\}, +), (\{1\}, \times)$
2	$C_2, (\mathbb{Z}_2, +_2)$	cyclic	$S^+(\square), (\mathbb{Z}_3^*, \times_3)$
3	$C_3, (\mathbb{Z}_3, +_3)$	cyclic	$S^+(\triangle), (\{0, 4, 8\}, +_{12}), (\{1, 4, 7\}, \times_9)$
4	$C_4, (\mathbb{Z}_4, +_4)$	cyclic	$(\mathbb{Z}_5^*, \times_5), S^+(\square), S(\triangle), (\{0, 3, 6, 9\}, +_{12}), (\{1, -1, i, -i\}, \times)$
$V, S(\square)$		abelian,	$(U_8, \times_8), (U_{12}, \times_{12}),$
		non-cyclic	$(\{1, 7, 9, 15\}, \times_{16}), (\{1, 9, 11, 19\}, \times_{20})$
5	$C_5, (\mathbb{Z}_5, +_5)$	cyclic	$S^+(\diamond)$
6	$C_6, (\mathbb{Z}_6, +_6)$	cyclic	$S^+(\square), (\mathbb{Z}_7^*, \times_7), (U_9, \times_9), (\{0, 2, 4, 6, 8, 10\}, +_{12}), (U_{14}, \times_{14})$
$S(\triangle)$		non-abelian	$S_3, \{e, (2 3), (2 4), (3 4), (2 3 4), (2 4 3)\}$
7	$C_7, (\mathbb{Z}_7, +_7)$	cyclic	$S^+(\text{heptagon})$
8	$C_8, (\mathbb{Z}_8, +_8)$	cyclic	$S^+(\text{octagon})$
$S(\text{cuboid})$		abelian	
(U_{15}, \times_{15})		abelian	(U_{20}, \times_{20})
$S(\square)$		non-abelian	
Q_8		non-abelian	

Proposition 17.2 (Useful results)

Let G be a group of finite order:

- (i) if each element except the identity has order 2, then G is abelian.
- (ii) if $\text{ord}(G) > 2$ and each element except e has order 2, then $4|\text{ord}(G)$.
- (iii) if $\text{ord}(G)$ is even, then at least one element of G has order 2.

Proof.

- (i) let $x, y \in G$, then xy is either the identity element or it has order 2 so that:

$$(xy)^2 = e \implies xyxy = e \implies xey = x^2yxy^2 \implies xy = yx \quad (16.2.1)$$

since $x^2 = y^2 = e$. The group G is therefore abelian.

- (ii) Firstly, by the previous point G must be abelian. Also, G has at least 3 elements, e, x, y , with $x^2 = y^2 = e$. Now consider $z = xy$, this must be distinct from x, y, e since

$$z = e \implies xy = e \implies y = x \quad (16.2.2)$$

$$z = x \implies xy = x \implies y = e \quad (16.2.3)$$

$$z = y \implies xy = y \implies x = e \quad (16.2.4)$$

It now remains to prove that $\{e, x, y, z\} \leq G$. We construct the following Cayley table:

	e	x	y	z
e	e	x	y	z
x	x	e	z	y
y	y	z	e	x
z	z	y	x	e

where for example $yxy = yyx = x$. The subgroup properties are then readily verified. Closure holds since every element in the body of the table is in $\{e, x, y, z\}$. The identity element of G is $e \in \{e, x, y, z\}$. Finally, all elements are self-inverse, and consequently their inverses belong to the same set.

We conclude that $\{e, x, y, z\} \leq G$, and by Lagrange's theorem, $4|\text{ord}(G)$.

- (iii) the elements that are not-self inverse can be paired up with their inverses, so they must be even. It follows that the number of self-inverse elements must also be even (for if they were odd then G would have odd order). The identity element is one such self-inverse element, so there must be at least one more self-inverse element, which of course has order 2.

■

Groups of order 1,2,3,5,7

Obviously, there is only one isomorphism class for groups of order 1, and that is C_1 .

For the other groups of prime order p , we have from the Corollary to Lagrange's theorem that they are isomorphic to $(\mathbb{Z}_p, +_p)$, and therefore belong to the same isomorphism class.

So the isomorphism class for each group of prime order p is the one containing C_p .

Groups of order 4

If G is a group of order 4, then by the Corollary to Lagrange's theorem, we must have that each element $g \in G$ must have order 1,2 or 4.

G has an element of group 4

If G has an element of order 4, then G is cyclic (generated by this element) and so isomorphic to C_4 . They all belong to the same isomorphism class.

G has no element of group 4

Only the identity element has order 1, so the other three elements must have order 2. By proposition 17.2 then, G is abelian, and if we let $G = \{e, x, y, z\}$, then by the same logic as in the proof of (ii) $z = xy$ and we retrieve the following Cayley table: which is the table

	e	x	y	z
e	e	x	y	z
x	x	e	z	y
y	y	z	e	x
z	z	y	x	e

of the Klein four-group V , so that $G \cong V$.

Therefore the two isomorphism classes for groups of order 4 are the one containing C_4 and the one containing V .

Groups of order 6

Suppose that G is a group of order 6, so that each element of G has order 1,2,3 or 6.

G has an element of group 6 In this case, G is a cyclic group, and is therefore isomorphic to C_6 . It can be classified in the isomorphism class containing C_6 .

G has no element of group 6 In this case, each non-identity element has order 2 or 3. We can assert that it must have at least one element of order 2 by proposition 17.2, and similarly there must be at least one element of order 3, for if they were all of order 2 then $4|ord(G)$ which clearly isn't the case.

So let $g, h \in G$ be some elements of order 2 and 3 respectively. We define:

$$H = \langle h \rangle = \{e, h, h^2\} \quad (16.2.5)$$

Obviously $g \neq H$, since all elements of H must have order 1 or 3 by the Corollary to Lagrange's theorem. We can then adopt the proof we used for Lagrange's theorem and write the following matrix:

$$\begin{pmatrix} e & h & h^2 \\ g & gh & gh^2 \end{pmatrix} \quad (16.2.6)$$

which contains six distinct elements, and must therefore include all elements of G .

Hence, we know that $G = \{e, h, h^2, g, g^2, gh, gh^2\}$. We construct the following incomplete Cayley table:

	e	h	h^2	g	gh	gh^2
e	e	h	h^2	g	gh	gh^2
h	h	h^2	e			
h^2	h^2	e	h			
g	g	gh	gh^2	e	h	h^2
gh	gh	gh^2	g			
gh^2	gh^2	g	gh			

To evaluate the missing entries, we need to calculate hg , which must be equal to gh or gh^2 (not g since it already appears in the same column). However, if $hg = gh$ then:

$$hg = gh \neq e \implies (hg)^3 = (hg)(gh)(hg) = g \neq e \quad (16.2.7)$$

so that hg has order greater than 3. The only possible value for $\text{ord}(hg)$ is then 6, a contradiction.

Therefore we must have that $hg = gh^2$. We then obtain (using the fact that each element must repeat only once in each row and column):

	e	h	h^2	g	gh	gh^2
e	e	h	h^2	g	gh	gh^2
h	h	h^2	e	gh^2	g	gh
h^2	h^2	e	h	gh	gh^2	g
g	g	gh	gh^2	e	h	h^2
gh	gh	gh^2	g	h^2	e	h
gh^2	gh^2	g	gh	h	h^2	e

This Cayley table is identical in structure to the tables of the groups S_3 and $S(\Delta)$. This means that the second isomorphism class is that containing $S(\Delta)$.

Therefore, the two isomorphism classes for groups of order 6 are the one containing C_6 and the one containing S_3 .

Unit E1: Cosets and normal subgroups

17.1 Matrix groups

We will introduce some important sets of matrices which form groups and subgroups under matrix multiplication. They are especially important in physics, we will study them in much more detail when studying representation theory and Lie groups.

Definition 18.1 (General linear group $GL(n, \mathbb{R})$)

Let $M(n, \mathbb{R})$ be the set of all real invertible $n \times n$ matrices. These form a group under matrix multiplication, called the **general linear group** denoted $GL(n, \mathbb{R})$.

Proof. We need to show that the group axioms are satisfied.

Closure Let $A, B \in GL(n, \mathbb{R})$, and let A^{-1} and B^{-1} be their inverses. Then, since $\det A \neq 0$ and $B \neq 0$, we find that:

$$\det(AB) = \det A \cdot \det B \neq 0 \implies AB \in GL(n, \mathbb{R}) \quad (17.1.1)$$

Associativity matrix multiplication is associative.

Identity The identity matrix \mathbb{I} is the identity of $GL(n, \mathbb{R})$. Indeed $\forall A \in GL(n, \mathbb{R})$:

$$\mathbb{I}A = A\mathbb{I} = A \quad (17.1.2)$$

as desired. Moreover, $\mathbb{I} \in GL(n, \mathbb{R})$, since $\det \mathbb{I} = 1 \neq 0$.

Inverses Let $A \in GL(n, \mathbb{R})$, then it must have an inverse A^{-1} , such that:

$$AA^{-1} = A^{-1}A = \mathbb{I} \quad (17.1.3)$$

Moreover, A^{-1} is invertible (its inverse if A , so $A^{-1} \in GL(n, \mathbb{R})$).

Therefore all the group axioms are satisfied, and $GL(n, \mathbb{R})$ form a group under matrix multiplication. ■

Clearly, $\dim \mathrm{GL}(n, \mathbb{R}) = n^2$. Indeed, the restriction that $\det \mathbf{A} \neq 0$ only restrict the values that the matrix elements cannot take, but it does not force some matrix elements to take a specific value. Hence they are spanned by the standard basis of $\mathrm{Mat}_n(\mathbb{R})$.

Definition 18.2 (Special linear group $SL(n, \mathbb{R})$)

The subset of $\mathrm{GL}(n, \mathbb{R})$ comprising of all invertible $n \times n$ matrices with unit determinant forms a subgroup of the general linear group, called the **special linear group**, and denoted $SL(n, \mathbb{R})$.

Proof. We need to show that the subgroup axioms are satisfied.

Closure Let $\mathbf{A}, \mathbf{B} \in SL(n, \mathbb{R})$, and let \mathbf{A}^{-1} and \mathbf{B}^{-1} be their inverses. Then, since $\det \mathbf{A} = 1$ and $\det \mathbf{B} = 1$, we find that:

$$\det(\mathbf{AB}) = \det \mathbf{A} \cdot \det \mathbf{B} = 1 \implies \mathbf{AB} \in SL(n, \mathbb{R}) \quad (17.1.4)$$

Identity The identity matrix $\mathbb{I} \in SL(n, \mathbb{R})$ since $\det \mathbb{I} = 1$.

Inverses Let $\mathbf{A} \in SL(n, \mathbb{R})$, then it must have an inverse \mathbf{A}^{-1} , and:

$$\det \mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}} = 1 \quad (17.1.5)$$

so $\mathbf{A}^{-1} \in SL(n, \mathbb{R})$.

Therefore all the subgroup axioms are satisfied, and $SL(n, \mathbb{R})$ forms a subgroup under matrix multiplication. ■

Perhaps more difficult to see is the fact that $\dim SL(n, \mathbb{R})$. Indeed, the most general form of a matrix $\mathbf{A} \in SL(n, \mathbb{R})$:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{pmatrix} \quad (17.1.6)$$

The condition $\det \mathbf{A} = 1$ can be expanded as:

$$\det \mathbf{A} = \sum_{\sigma \in S_n} \mathrm{sgn}(\sigma) a_{\sigma(1),1} a_{\sigma(2),2} a_{\sigma(3),3} \dots a_{\sigma(n),n} \quad (17.1.7)$$

We can solve this equation for a_{nn} in terms of the other $n^2 - 1$ components, so there are in total $n^2 - 1$ independent elements in \mathbf{A} , giving:

$$\dim SL(n, \mathbb{R}) = n^2 - 1 \quad (17.1.8)$$

The extension of these results for $\mathrm{SL}(n, \mathbb{C})$ is immediate:

$$\dim \mathrm{SL}(n, \mathbb{C}) = 2 \cdot \dim \mathrm{SL}(n, \mathbb{R}) = 2n^2 - 2 \quad (17.1.9)$$

Definition 18.3 (Orthogonal group $O(n, \mathbb{R})$)

The subset of $\mathrm{GL}(n, \mathbb{R})$ comprising of all invertible $n \times n$ orthogonal matrices ($A^T A = \mathbb{I}$) forms a subgroup of the general linear group, called the **orthogonal group**, and denoted $O(n, \mathbb{R})$.

Proof. Firstly note that if $A \in O(n, \mathbb{R})$, then:

$$\det AA^T = (\det\{A\})^2 = \det \mathbb{I} = 1 \implies \det\{A\} = \pm 1 \quad (17.1.10)$$

so A must also be invertible.

We need to show that the subgroup axioms are satisfied.

Closure Let $A, B \in O(n, \mathbb{R})$. Then we find that:

$$(AB)(AB)^T = ABB^T A^T = \mathbb{I} \quad (17.1.11)$$

so $AB \in O(n, \mathbb{R})$.

Identity The identity matrix $\mathbb{I} \in O(n, \mathbb{R})$ since $\mathbb{I}^T = \mathbb{I}$.

Inverses Let $A \in O(n, \mathbb{R})$. Then its inverse A^{-1} in $\mathrm{GL}(n, \mathbb{R})$ satisfies

$$(A^{-1})(A^{-1})^T = (A^{-1})(A^T)^{-1} = (A^T A)^{-1} = \mathbb{I}^{-1} = \mathbb{I} \quad (17.1.12)$$

so that $A \in O(n, \mathbb{R})$ as desired. So $A^{-1} \in \mathrm{SL}(n, \mathbb{R})$.

Therefore all the subgroup axioms are satisfied, and $O(n, \mathbb{R})$ forms a group under matrix multiplication. ■

Again, finding the dimension of the orthogonal group of order n is slightly more involved. Consider the equation:

$$AA^T - \mathbb{I} = 0 \quad (17.1.13)$$

Note that $(AA^T)^T = (A^T)^T A^T = AA^T$, so AA^T only has $\frac{1}{2}n(n+1)$ independent components, namely the n diagonal components, and then the lower/upper triangular components $\frac{1}{2}(n^2 - n)$.

n). Hence can be expressed as:

$$\begin{pmatrix} b_{11} & \dots & \dots & B \\ \vdots & b_{22} & & \vdots \\ \vdots & & & \vdots \\ B & \dots & & b_{nn} \end{pmatrix} = 0 \quad (17.1.14)$$

where we absorbed \mathbb{I} into the diagonal b_{kk} components. As we said earlier the above matrix equation has $\frac{1}{2}n(n+1)$ independent equations in a_{ij} , fixing $\frac{1}{2}n(n+1)$ elements of A . Hence:

$$\dim O(n, \mathbb{R}) = n^2 - \frac{1}{2}n(n+1) = \frac{1}{2}n(n-1) \quad (17.1.15)$$

Definition 18.4 (Special orthogonal group $SU(n, \mathbb{R})$)

The subset of $O(n, \mathbb{R})$ comprising of all invertible $n \times n$ orthogonal matrices with unit determinant forms a subgroup of the orthogonal group, called the **special orthogonal group**, and denoted $SU(n, \mathbb{R})$.

The proof is a combination of the proofs in definitions 18.2 and 18.3.

The dimension of $SU(n, \mathbb{R})$ is surprisingly equal to the dimension of $O(n, \mathbb{R})$, despite the additional constraint that $\det A = +1$. Note that for $A = O(n, \mathbb{R})$, we have $A = \pm 1$, so we're only removing the matrices with $\det A = -1$. The constraints however remain the same, so we still have $\frac{1}{2}n(n-1)$ constraints.

Definition 18.4 (Unitary group $U(n, \mathbb{C})$)

The subset of $GL(n, \mathbb{C})$ comprising of all unitary $n \times n$ matrices ($AA^\dagger = \mathbb{I}$) forms a subgroup of the general linear group, called the **unitary group**, and denoted $U(n, \mathbb{C})$.

Proof. We need to show that the subgroup axioms are satisfied.

Closure Let $A, B \in U(n, \mathbb{C})$. Then we find that:

$$(AB)(AB)^\dagger = ABB^\dagger A^\dagger = \mathbb{I} \quad (17.1.16)$$

so $AB \in O(n, \mathbb{C})$.

Identity The identity matrix $\mathbb{I} \in U(n, \mathbb{C})$ since $\mathbb{I}\mathbb{I}^\dagger = \mathbb{I}$.

Inverses Let $A \in U(n, \mathbb{C})$. Then its inverse A^{-1} in $GL(n, \mathbb{C})$ satisfies

$$(A^{-1})(A^{-1})^\dagger = (A^{-1})(A^\dagger)^{-1} = (A^\dagger A)^{-1} = \mathbb{I}^{-1} = \mathbb{I} \quad (17.1.17)$$

so that $A^{-1} \in U(n, \mathbb{C})$ as desired.

Therefore all the subgroup axioms are satisfied, and $U(n, \mathbb{C})$ forms a group under matrix multiplication. ■

What is the dimension of $U(n, \mathbb{C})$? Note that $\text{Mat}_n(\mathbb{C})$ has dimension $2n^2$, n^2 free real parameters, and n^2 free complex parameters. Now let's see how many constraints unitarity ($\mathbf{A}\mathbf{A}^\dagger = \mathbb{I}$) imposes.

Note that $\mathbf{A}\mathbf{A}^\dagger$ is hermitian, since $(\mathbf{A}\mathbf{A}^\dagger)^\dagger = (\mathbf{A}^\dagger)^\dagger \mathbf{A}^\dagger = \mathbf{A}\mathbf{A}^\dagger$. A hermitian matrix has a total of n^2 free parameters, n diagonal (the diagonal components have to be real, since if they had an imaginary part, the hermitian conjugate would turn it negative, thus violating hermiticity), and $n^2 - n$ off-diagonal ($\frac{1}{2}(n^2 - n)$ for the real part, and $\frac{1}{2}(n^2 - n)$ for the imaginary part). Therefore the hermiticity constraint consists of n^2 independent equations, and thus fixes n^2 elements. Hence:

$$\dim U(n, \mathbb{C}) = 2n^2 - n^2 = n^2 \quad (17.1.18)$$

Definition 18.4 (Special unitary group $SU(n, \mathbb{C})$)

The subset of $U(n, \mathbb{C})$ comprising of all invertible $n \times n$ hermitian matrices with unit determinant forms a subgroup of the unitary group, called the **special unitary group**, and denoted $SU(n, \mathbb{C})$.

Note that the determinant of $\mathbf{A} \in U(n, \mathbb{R})$ is such that:

$$\det \mathbf{A}\mathbf{A}^\dagger = (\det \mathbf{A})(\det \mathbf{A})^* = |\det \mathbf{A}|^2 = 1 \implies \det \mathbf{A} = e^{i\theta}, \forall \theta \in [0, 2\pi] \quad (17.1.19)$$

whereas we are restricting $\det \mathbf{A} = 1 \implies \theta = 0$. Unlike in the orthogonal and special orthogonal group case, where we were simply restricting the sign of the determinant, here we are restricting a whole continuum of values that the determinant can take, hence we have an additional constraint. Therefore:

$$\dim SU(n, \mathbb{C}) = n^2 - 1 \quad (17.1.20)$$

We summarize the dimensionalities of these matrix groups below:

Theorem 18.5 (Dimensions of matrix groups)

We have that:

- (i) $\dim GL(n, \mathbb{R}) = n^2$
- (ii) $\dim SL(n, \mathbb{R}) = n^2 - 1$
- (iii) $\dim O(n, \mathbb{R}) = \frac{1}{2}n(n - 1)$
- (iv) $\dim SU(n, \mathbb{R}) = \frac{1}{2}n(n - 1)$
- (v) $\dim U(n, \mathbb{C}) = n^2$
- (vi) $\dim SU(n, \mathbb{C}) = n^2 - 1$

17.2 Cosets

Definition 18.6 (Left coset)

Let $H < G$ be a subgroup of G , and let $g \in G$. Then, the **left coset** gH of H in g is given by:

$$gH = \{gh : h \in H\} \quad (17.2.1)$$

and is the subset of G obtained by composing each element in H with g to the left.

Example. For example, let's try to find all the left cosets of the subgroup $H = \{e, s\}$ in the group $S(\Delta)$.

We can use the Cayley table for $S(\Delta)$:

$$eH = e\{e, s\} = \{e, s\} \quad (17.2.2)$$

$$aH = a\{e, s\} = \{a, r\} \quad (17.2.3)$$

$$bH = b\{e, s\} = \{b, t\} \quad (17.2.4)$$

$$rH = r\{e, s\} = \{r, a\} = aH \quad (17.2.5)$$

$$sH = s\{e, s\} = \{s, e\} = eH \quad (17.2.6)$$

$$tH = t\{e, s\} = \{t, b\} = bH \quad (17.2.7)$$

so the distinct left cosets are $\{e, s\}, \{a, r\}, \{b, t\}$. ◀

Example. For example, let's try to find all the left cosets of the subgroup $H = \{1, 2, 4\}$ in \mathbb{Z}_7^* .

We find that:

$$1H = 1\{1, 2, 4\} = \{1, 2, 4\} \quad (17.2.8)$$

$$2H = 2\{1, 2, 4\} = \{2, 4, 1\} \quad (17.2.9)$$

$$3H = 3\{1, 2, 4\} = \{3, 6, 5\} \quad (17.2.10)$$

$$4H = 4\{1, 2, 4\} = \{4, 1, 2\} \quad (17.2.11)$$

$$5H = 5\{1, 2, 4\} = \{5, 3, 6\} \quad (17.2.12)$$

$$6H = 6\{1, 2, 4\} = \{6, 5, 3\} \quad (17.2.13)$$

so we see that the distinct left cosets are $\{1, 2, 4\}$ and $\{3, 5, 6\}$. ◀

It is interesting to note that the distinct left cosets of $\{e, s\}$ in $S(\Delta)$ and the distinct left cosets of $\{1, 2, 4\}$ in \mathbb{Z}_7^* partition their respective groups.

For example $\{e, s\} \cup \{a, r\} \cup \{b, t\} = S(\Delta)$, and all three sets are disjoint.

This is not a coincidence, it turns out that all distinct left cosets are partitions.

Theorem 18.7 (Left coset partition)

Let $H < G$ be a subgroup of a group G . Then the distinct left cosets of H in G form a partition of G .

Proof. Recall that the equivalence classes of an equivalence relation on some set X form a partition of the set X . Consequently, if we can show that a particular operation has left cosets of a subgroup H in a group G , then immediately we find that the left cosets form a partition.

Lemma. Let \sim be the relation defined on G by:

$$x \sim y \text{ if } x \in yH \quad (17.2.14)$$

Then \sim is an equivalence relation. Indeed:

- (i) **Reflexive property:** let $x \in G$, we have to show that $x \sim x$, that is, $x \in xH$. This is clearly true since $x = xe$, and $e \in H$ due to the subgroup axioms.
- (ii) **Symmetric property:** let $x, y \in G$, and let $x \sim y$, so that $x \in yH$. Therefore, $\exists h \in H$ such that $x = yh \implies y = xh^{-1}$. Now, due to the inverses property of subgroups, $h^{-1} \in H \implies y \in xH$ so $y \sim x$.
- (iii) **Transitive property:** let $x, y, z \in G$, and suppose $x \sim y, y \sim z$, so $x \in yH$ and $y \in zH$. Therefore, $\exists h_1, h_2 \in H$ such that:

$$x = yh_1 \text{ and } y = zh_2 \implies x = zh_1h_2 \quad (17.2.15)$$

Since $h_1h_2 \in H$, we find that $x \in zH$, and thus $x \sim z$.

We have therefore shown that \sim is an equivalence relation. Each element $x \in G$ has equivalence class:

$$[x] = \{y \in G : y \sim x\} = \{y \in G : y \in xH\} = xH \quad (17.2.16)$$

which are the left cosets of H in G . It follows then that the left cosets (which are the equivalence classes) form a partition of G . ■

Proposition 18.9 (Properties of left cosets) Let $H < G$, then:

- (i) $\forall g \in G, g \in gH$
- (ii) one of the left cosets of H in G is H
- (iii) Any two left cosets g_1H and g_2H are either the same or disjoint
- (iv) If $|H| < \infty$, then each left coset gH has the same number of elements.

Proof. (i) Trivial, since $e \in H$ due to subgroup axioms.

(ii) $H = \{h : h \in H\} = \{eh : h \in H\} = eH$ so H is indeed a left coset.

(iii) Immediate from theorem 18.8.

(iv) Let $|H| = m$ so that $H = \{h_1, h_2, \dots, h_m\}$, and let $g \in G$. Then:

$$gH = \{gh_1, gh_2, \dots, gh_m\} \quad (17.2.17)$$

Suppose $gh_i = gh_j$, then by the left cancellation law $h_i = h_j$, so it follows that $|gH| = |H| = m$.

■

Example. Let's try to find all the left cosets of $H = \{e, a, b, c\}$ in $S(\square)$. The remaining elements are r, s, t, u , and the remaining left cosets all contain 4 elements. So the only way to make them disjoint is if we have $\{r, s, t, u\}$ as the other coset. Hence the distinct left cosets are $\{e, a, b, c\}$ and $\{r, s, t, u\}$.

To partition a finite group G into left cosets of some subgroup $H < G$:

- (i) H is the first cosets
- (ii) find an element in G that is not in H , and determine gH .
- (iii) repeat until all elements in G have been exhausted.

◀

Example. Consider the group $U_{20} = \{1, 3, 7, 9, 11, 13, 17\}$ and $H = \{1, 19\}$.

Firstly, H is one of the left cosets of H in U_{20} , so $\{1, 19\}$. One remaining element is 3, and its corresponding left coset is:

$$3H = 3\{1, 19\} = \{3, 17\} \quad (17.2.18)$$

An element missing from both H and $3H$ is 7, and its corresponding left coset is:

$$7H = 7\{1, 19\} = \{7, 13\} \quad (17.2.19)$$

The final missing element from all these left cosets is 9:

$$9H = 9\{1, 19\} = \{9, 11\} \quad (17.2.20)$$

So we have found that the distinct left cosets are $\{1, 19\}, \{3, 17\}, \{7, 13\}$ and $\{9, 11\}$.

◀

Example. Let's try to partition the alternating group of order 4:

$$A_4 = \{e, (1 2)(3 4), (1 3)(2 4), (1 4)(2 3), (1 2 3), (1 3 2), (1 2 4), (1 4 2), \quad (17.2.21)$$

$$(1 3 4), (1 4 3), (2 3 4), (2 4 3)\} \quad (17.2.22)$$

into left cosets of the subgroup $H = \{e, (1 2)(3 4), (1 3)(2 4), (1 4)(2 3)\}$.

Firstly, we note that H itself is a left coset in A_4 . A remaining element not included is for example $(1 2 3)$, and its associated left coset is:

$$(1 2 3)H = \{(1 2 3), (1 3 4), (2 4 3), (1 4 2)\} \quad (17.2.23)$$

Therefore, the only remaining permutations are $(1 3 2), (1 2 4), (1 4 3), (2 3 4)$. Since the left cosets must contain as many elements as H , that is 4, and they must be disjoint, there is only one remaining left coset $\{(1 3 2), (1 2 4), (1 4 3), (2 3 4)\}$. Hence the partition of A_4 into the left cosets of H is:

$$H \cup \{(1 3 2), (1 2 4), (1 4 3), (2 3 4)\} = A_4 \quad (17.2.24)$$

◀

We can use the left cosets to provide a more efficient prove of Lagrange's theorem.

Theorem 17.1 (Lagrange's Theorem)

Let G be a finite group, and let $H \leq G$. Then $\text{ord}(H)|\text{ord}(G)$, that is, the order of H divides the order of G .

Proof. Let $|G| = n$ and $|H| = m$, and let the number of left cosets of H in G be k . Since each left coset has m elements, and they partition G into disjoint sets, it follows that $n = m \cdot k$, that is, $|H|$ divides $|G|$. ■

17.3 Right cosets

Definition 18.10 (Right cosets)

Let $H < G$ be a subgroup of G , and let $g \in G$. Then, the **right coset** Hg of H in g is given by:

$$Hg = \{hg : h \in H\} \quad (17.3.1)$$

and is the subset of G obtained by composing each element of H with g to the right.

Example. Let's find all the right cosets of $H = \{e, r\}$ in $S(\square)$.

We find that:

$$He = \{e, r\}e = \{e, r\} \quad (17.3.2)$$

$$Ha = \{e, r\}a = \{a, u\} \quad (17.3.3)$$

$$Hb = \{e, r\}b = \{b, t\} \quad (17.3.4)$$

$$Hc = \{e, r\}c = \{c, s\} \quad (17.3.5)$$

$$Hr = \{e, r\}r = \{r, e\} \quad (17.3.6)$$

$$Hs = \{e, r\}s = \{s, c\} \quad (17.3.7)$$

$$Ht = \{e, r\}t = \{t, b\} \quad (17.3.8)$$

$$Hu = \{e, r\}u = \{u, a\} \quad (17.3.9)$$

so the distinct right cosets are: $\{e, r\}, \{a, u\}, \{b, t\}, \{c, s\}$. ◀

It is interesting to note that the right coset of H in some element g is not necessarily equal to the left coset of H in g .

All the results proven for left cosets are easily proven for right cosets as well.

Theorem 18.11 (Right coset partition)

Let $H < G$ be a subgroup of G , then the distinct right cosets of H in G form a partition of G .

Proposition 18.12 (Properties of right coset)

Let $H < G$, then:

- (i) $g \in Hg$, for all $g \in G$
- (ii) H is a right coset of H in G
- (iii) two right cosets Hg_1 and Hg_2 are either the same set or disjoint
- (iv) if $|H| < \infty$ then $\forall g \in G |Hg| = |H|$.

Example. Let's try to partition $S(\Delta)$ into right cosets of $H = \{e, s\}$.

Firstly, we note that all right cosets must have 2 elements, and one of these right cosets is H itself. An element in G that does not belong to H is a , and its right coset is:

$$Ha = \{e, s\}a = \{a, t\} \quad (17.3.10)$$

Now another element in G that does not belong to these two cosets is b , so:

$$Hb = \{e, s\}b = \{b, r\} \quad (17.3.11)$$

We have exhausted all elements of $S(\Delta)$, so we may write the partition of the latter

as:

$$S(\Delta) = \{e, s\} \cup \{a, t\} \cup \{b, r\} \quad (17.3.12)$$

◀

Note that if G is abelian and $H < G$, then H is also abelian and thus:

$$Hg = \{hg : \forall h \in H\} = \{gh : \forall h \in H\} = gH \quad (17.3.13)$$

so the left and right cosets are the same.

Theorem 18.13 (Transforming left coset to right coset)

Let $H < G$. Then if every element in the partition of G into left cosets of H is replaced by its inverse, the result is the partition of G into right cosets of H (and viceversa).

Proof. Consider a pair of elements x, y in the same left coset of H . Therefore,

Suppose $x \in gH$ for some $g \in G$. Then, we need to prove that $x^{-1} \in Hg^{-1}$ for some $g' \in G$. Indeed:

$$gH = \{gh : h \in H\} \text{ and } \{(gh)^{-1} : h \in H\} = \{h^{-1}g^{-1} : h \in H\} = \{hg^{-1} : h \in H\} = Hg^{-1} \quad (17.3.14)$$

due to the inverse axiom of subgroups¹. It follows that if we replace every element in gH by its inverse, we get a right coset Hg^{-1} . ■

Immediately, we find that since the act of replacing each element by its inverse is bijective, the number of distinct left and right cosets is the same.

Proposition 18.14 (Number of left and right cosets) Let $H < G$, then the number of distinct left cosets of H in G is equal to the number of distinct right cosets of H in G .

Definition 18.15 (Index) Let $H < G$. The **index** of H in G is the number of distinct left cosets (or equivalently distinct right cosets) of H in G .

For finite groups, we have a nice expression for the index.

¹indeed suppose

$$H = \{h_1, h_2, \dots, h_n, h_1^{-1}, h_2^{-1}, \dots, h_n^{-1}\}$$

then

$$\{h^{-1} : h \in H\} = \{h_1^{-1}, h_2^{-1}, \dots, h_n^{-1}, h_1, h_2, \dots, h_n\} = H$$

Proposition 18.16 (Finite index)

Let $H < G$, then the index of H in G is $\frac{|G|}{|H|}$.

Proof. The left cosets of H partition G , and since each left coset has $|H|$ elements, the number of left cosets (the index) must be $\frac{|G|}{|H|}$. \blacksquare

Example. Let's partition $(2\mathbb{Z}, +)$ into cosets of $6\mathbb{Z}$.

Note that $2\mathbb{Z} = \{\dots, -6, -4, -2, 0, 2, 4, 6, \dots\}$, and $6\mathbb{Z} = \{\dots, -18, -12, -6, 0, 6, 12, 18, \dots\}$. We know that $6\mathbb{Z} < 2\mathbb{Z}$ since $6\mathbb{Z}$ is generated by $6 \in 2\mathbb{Z}$, and is therefore a cyclic subgroup of $2\mathbb{Z}$.

Now note that $6\mathbb{Z}$ is itself a coset so:

$$0 + 6\mathbb{Z} = \{6k : k \in \mathbb{Z}\} \quad (17.3.15)$$

An element that was not included is 2:

$$2 + 6\mathbb{Z} = \{\dots, -16, -10, -4, 2, 8, 14, 20, \dots\} = \{2 + 6k : k \in \mathbb{Z}\} \quad (17.3.16)$$

An element that was not included is 4, so:

$$4 + 6\mathbb{Z} = \{\dots, -14, -8, -2, 4, 10, 16, 22, \dots\} = \{4 + 6k : k \in \mathbb{Z}\} \quad (17.3.17)$$

Note that:

$$\{2(3k+2) : k \in \mathbb{Z}\} \cup \{2(3k+1) : k \in \mathbb{Z}\} \cup \{2(3k) : k \in \mathbb{Z}\} = \{2k : k \in \mathbb{Z}\} = 2\mathbb{Z} \quad (17.3.18)$$

so we do indeed have a partition. Moreover the index of $6\mathbb{Z}$ in $2\mathbb{Z}$ is 3, something that we could not have found using proposition 18.16. \blacktriangleleft

17.4 Normal subgroups

Although this doesn't generally happen, for certain special groups the left coset partition and right coset partition can be exactly the same.

Definition 18.17 (Normal subgroup)

Let G be a group and let $H < G$. We say that H is a **normal subgroup** if the left coset partition of G in H and the right coset partition of G in H are the same. We say that H is normal in G as well, denoted as $N \trianglelefteq G$.

Proposition 18.18 (Standard normal subgroups) For any group G :

- (i) the trivial subgroup $\{e\}$
- (ii) G

are both normal subgroups.

Proof. In the first case the left coset partition contains one-element subsets of G , and so does the right coset partition. In the second case the left coset partition is simply G , and so is the right coset partition. ■

Example. Consider A_4 and $H = \{e, (1 2)(3 4), (1 3)(2 4), (1 4)(2 3)\}$. The partition of A_4 into the left cosets of H was found previously to be:

$$\{e, (1 2)(3 4), (1 3)(2 4), (1 4)(2 3)\} \cup \{(1 3 2), (1 2 4), (1 4 3), (2 3 4)\} = A_4 \quad (17.4.1)$$

Now let's try to find the right coset partition. We can do this by simply replacing each element in the left coset partition by its inverse:

$$\{e, (1 2)(3 4), (1 3)(2 4), (1 4)(2 3)\} \cup \{(1 2 3), (1 4 2), (1 3 4), (2 4 3)\} = A_4 \quad (17.4.2)$$

Note that the left and right coset partition are identical, hence H is indeed a normal subgroup of G . ◀

Theorem 18.19 (Normal subgroups of abelian groups) Every subgroup H of an abelian group G is normal.

Proof. Let $H < G$, and let $g \in G$. Note that:

$$gH = \{gh : h \in H\} = \{hg : h \in H\} = Hg \quad (17.4.3)$$

thus H is normal. ■

Proposition 18.20 (Subgroups of index 2) Every subgroup of index 2 in a group is a normal subgroup

Proof. Let $H < G$ so that it has exactly two left cosets and exactly two right cosets in G . One of these cosets is H itself, so the other coset must be $G \setminus H$ both for the left and right coset partition. The two partitions are therefore identical, proving that H is a normal subgroup of G . ■

Example. If $n \geq 2$, it follows that A_n is a normal subgroup of S_n , since $\frac{|S_n|}{|A_n|} = 2$, so A_n is normal to S_n . If instead $n = 1$, then $A_1 = S_1$, and from proposition 18.18 we see that A_1 is normal to S_1 . \blacktriangleleft

Example. Consider $S(\square)$ and $S^+(\square)$. Since the number of direct symmetries is equal to the indirect symmetries, we find that the index of $S^+(\square)$ in $S(\square)$ is $\frac{|S(\square)|}{|S^+(\square)|} = 2$, and thus $S^+(\square)$ is a normal subgroup of $S(\square)$. \blacktriangleleft

Proposition 18.21 (Equivalent condition for normality)

Let $H < g$, then H is normal in G iff $gH = Hg$, $\forall g \in G$.

Proof. \implies Suppose that H is normal in G , and let $g \in G$. Then $g \in gH$ and $g \in Hg$.

Now since the left and right coset partitions are identical, and the cosets are disjoint, we must have that since g belongs to both Hg and gH , $Hg = gH$ as desired.

\impliedby Suppose $gH = Hg$, then it follows that the left coset and right coset partitions are the same. \blacksquare

Unit E2: Quotient groups and conjugacy

18.1 Quotient groups

Definition 19.1 (Set composition) Let G be a group, then the operation \cdot , called a **set composition** in G , is defined as:

$$X \cdot Y = \{xy : x \in X, y \in Y\} \quad (18.1.1)$$

where $X, Y \subseteq G$.

Note that for an arbitrary group, set composition is not necessarily commutative. Only for Abelian groups is set composition is commutative.

Consider the following Cayley table for the cosets of the normal subgroup $\{e, b\}$ in $S(\square)$:

\cdot	$\{e, b\}$	$\{a, c\}$	$\{r, t\}$	$\{s, u\}$
$\{e, b\}$	$\{e, b\}$	$\{a, c\}$	$\{r, t\}$	$\{s, u\}$
$\{a, c\}$	$\{a, c\}$	$\{e, b\}$	$\{s, u\}$	$\{r, t\}$
$\{r, t\}$	$\{r, t\}$	$\{u, s\}$	$\{e, b\}$	$\{a, c\}$
$\{s, u\}$	$\{s, u\}$	$\{r, t\}$	$\{a, c\}$	$\{e, b\}$

Interestingly all the sets in the body of the table are also cosets of $\{e, b\}$. Therefore the set of these cosets are closed under set composition. It turns out that we can extend this result more generally for any cosets of a normal subgroup N of a group G .

Theorem 19.2 (Closure of cosets under set composition) Let $N \trianglelefteq G$, then:

$$xN \cdot yN = (xy)N, \forall x, y \in G \quad (18.1.2)$$

Proof. Let us firstly show that $xN \cdot yN \subseteq (xy)N$. Indeed, let $z \in xN \cdot yN$, so that there are some $n_1, n_2 \in N$ such that:

$$z = xn_1yn_2 \quad (18.1.3)$$

Since N is a normal subgroup, $n_1y \in Ny \implies n_1y \in yN$ so

$$n_1y = yn_3, n_3 \in N \quad (18.1.4)$$

Then:

$$z = xyn_3n_2 = xyn, n = n_3n_2 \in N \quad (18.1.5)$$

implying that $z \in xyN$, as desired.

Let us now show that $(xy)N \subseteq xN \cdot yN$. Suppose that $z \in (xy)N$, so that there is some $n_1 \in N$ such that:

$$z = xyn_1 \quad (18.1.6)$$

Since $x \in xN$ and $yn_1 \in yN$, we find that $z \in xN \cdot yN$, as desired. ■

Due to this theorem, we see that if we perform set composition on the cosets of some normal subgroup $N \trianglelefteq G$ in G , then the result will be another coset.

If we examine the Cayley table for the cosets of $\{e, b\}$ in $S(\square)$, we find that not only is closure under \cdot satisfied, all the other group properties are also satisfied! Associativity of set composition follows from associativity of composition in $S(\square)$. Moreover, the identity element can be verified to be $\{e, b\}$ so that all cosets are self-inverse.

Let us prove that the cosets of a normal subgroup form a group under set composition in the most general case.

Theorem 19.3 (Group of cosets) Let $N \trianglelefteq G$, then the set of cosets of N in G forms a group under set composition. This group is called the **quotient group** of G by N , denoted $G \setminus N$, and often read G mod N for brevity.

Proof.

Closure We have proven closure in Theorem 19.2

Associativity Let xN, yN, zN be cosets of N in G . Then:

$$xN \cdot (yN \cdot zN) = xN \cdot (yz)N = (x(yz))N = (xyz)N \quad (18.1.7)$$

since G is a group, and therefore its operation is associative. Similarly

$$(xN \cdot yN) \cdot zN = ((xy)z)N = (xyz)N \quad (18.1.8)$$

Identity We prove that eN is the identity element. Indeed:

$$eN \cdot xN = (ex)N = xN, \forall x \in G \quad (18.1.9)$$

and similarly:

$$xN \cdot eN = (xe)N = xN, \quad \forall x \in G \quad (18.1.10)$$

as desired.

Inverses Suppose xN is a coset of N in G . Then:

$$x^{-1}N \cdot xN = (x^{-1}x)N = eN, \quad \forall x \in G \quad (18.1.11)$$

and similarly:

$$xN \cdot x^{-1}N = (xx^{-1})N = eN, \quad \forall x \in G \quad (18.1.12)$$

Therefore, $x^{-1}N$ is the inverse of xN , and must belong to the set of cosets since $x \in G \implies x^{-1} \in G$

■

Note that if G is finite, then $|G \setminus N| = \frac{|G|}{|N|}$ is the number of cosets of N in G .

We can use our knowledge of normal subgroups and quotient groups to explain the "block" effect in the Cayley table of $S(\mathcal{F})$.

Indeed, note that since $S^+(\mathcal{F})$ is a normal subgroup, we can construct its cosets in $S(\mathcal{F})$. Since $S^+(\mathcal{F})$ has index 2, there will be two such cosets, with it being one of them. The remaining coset must therefore be the set of indirect symmetries $S^-(\mathcal{F})$. Consequently:

.	$S^+(\mathcal{F})$	$S^-(\mathcal{F})$
$S^+(\mathcal{F})$	$S^+(\mathcal{F})$	$S^-(\mathcal{F})$
$S^-(\mathcal{F})$	$S^-(\mathcal{F})$	$S^+(\mathcal{F})$

If we now expand $S^+(\mathcal{F})$ and $S^-(\mathcal{F})$ into its various components, then we find the blocks:

\circ	$e \ a \ b \ c \ r \ s \ t \ u$	\circ	direct	indirect
e	$e \ a \ b \ c \ r \ s \ t \ u$			
a	$a \ b \ c \ e \ s \ t \ u \ r$	direct	direct	indirect
b	$b \ c \ e \ a \ t \ u \ r \ s$			
c	$c \ e \ a \ b \ u \ r \ s \ t$			
r	$r \ u \ t \ s \ e \ c \ b \ a$			
s	$s \ r \ u \ t \ a \ e \ c \ b$			
t	$t \ s \ r \ u \ b \ a \ e \ c$	indirect	indirect	direct
u	$u \ t \ s \ r \ c \ b \ a \ e$			

Figure 18.1. Block effect in $S(\square)$

Example. Consider the subgroup $H = \langle 6 \rangle$ of \mathbb{Z}_{12} . The elements of H are:

$$\langle 6 \rangle = \{0, 6\} \implies \text{ord}(6) = 2 \quad (18.1.13)$$

Since \mathbb{Z}_{12} is an abelian group, all of its subgroups are normal, including $\langle 6 \rangle$, which is its cyclic subgroup of order 3.

The cosets of H in \mathbb{Z}_{12} must be:

$$H = \{0, 6\} \quad (18.1.14)$$

$$1 + H = \{1, 7\} \quad (18.1.15)$$

$$2 + H = \{2, 8\} \quad (18.1.16)$$

$$3 + H = \{3, 9\} \quad 4 + H = \{4, 10\} \quad (18.1.17)$$

$$5 + H = \{5, 11\} \quad (18.1.18)$$

The quotient group, formed by the above cosets, must then have Cayley table:

$+$	H	$1 + H$	$2 + H$	$3 + H$	$4 + H$	$5 + H$
H	H	$1 + H$	$2 + H$	$3 + H$	$4 + H$	$5 + H$
$1 + H$	$1 + H$	$2 + H$	$3 + H$	$4 + H$	$5 + H$	H
$2 + H$	$2 + H$	$3 + H$	$4 + H$	$5 + H$	H	$1 + H$
$3 + H$	$3 + H$	$4 + H$	$5 + H$	H	$1 + H$	$2 + H$
$4 + H$	$4 + H$	$5 + H$	H	$1 + H$	$2 + H$	$3 + H$
$5 + H$	$5 + H$	H	$1 + H$	$2 + H$	$3 + H$	$4 + H$

For example, $(4 + H) + (3 + H) = (4 +_{12} 3)H = 7 + H = 1 + H$. We clearly see that H is the identity element, and that $H, 3 + H$ are self inverse, whereas $2 + H$ and $4 + H$ are inverses of each other and so are $1 + H$ and $5 + H$.

We see that this Cayley table has the same structure as the sixth order cyclic group C_6 . ◀

18.2 Quotient group of infinite groups

Consider the group $\mathbb{Z} \setminus 4\mathbb{Z}$, that is, the set of cosets of $4\mathbb{Z} = \{4n : n \in \mathbb{Z}\}$ in \mathbb{Z} . Its elements are:

$$4\mathbb{Z} \quad (18.2.1)$$

$$1 + 4\mathbb{Z} = \{1 + 4n : n \in \mathbb{Z}\} \quad (18.2.2)$$

$$2 + 4\mathbb{Z} = \{2 + 4n : n \in \mathbb{Z}\} \quad (18.2.3)$$

$$3 + 4\mathbb{Z} = \{3 + 4n : n \in \mathbb{Z}\} \quad (18.2.4)$$

$$(18.2.5)$$

We know that these are all the cosets since they partition \mathbb{Z} . We may construct the Cayley table for $\mathbb{Z} \setminus 4\mathbb{Z}$ as:

+	4 \mathbb{Z}	1 + 4 \mathbb{Z}	2 + 4 \mathbb{Z}	3 + 4 \mathbb{Z}
4 \mathbb{Z}	4 \mathbb{Z}	1 + 4 \mathbb{Z}	2 + 4 \mathbb{Z}	3 + 4 \mathbb{Z}
1 + 4 \mathbb{Z}	2 + 4 \mathbb{Z}	3 + 4 \mathbb{Z}	4 \mathbb{Z}	
2 + 4 \mathbb{Z}	2 + 4 \mathbb{Z}	3 + 4 \mathbb{Z}	4 \mathbb{Z}	1 + 4 \mathbb{Z}
3 + 4 \mathbb{Z}	3 + 4 \mathbb{Z}	4 \mathbb{Z}	1 + 4 \mathbb{Z}	2 + 4 \mathbb{Z}

Note that this has the exact same structure as the Cayley table for \mathbb{Z}_4 , so $\mathbb{Z}_4 \cong \mathbb{Z} \setminus 4\mathbb{Z}$ through the isomorphism:

$$\phi : \mathbb{Z}/4\mathbb{Z} \rightarrow \mathbb{Z}_4 \quad (18.2.6)$$

$$a + 4\mathbb{Z} \mapsto a \forall a \in \mathbb{Z}_4 \quad (18.2.7)$$

We can prove this result more generally.

Proposition 19.4 ($\mathbb{Z} \setminus n\mathbb{Z} \cong \mathbb{Z}_n$)

For $n \geq 2$, then $\mathbb{Z} \setminus n\mathbb{Z} \cong \mathbb{Z}_n$. One isomorphism between them is:

$$\phi : \mathbb{Z}/n\mathbb{Z} \rightarrow \mathbb{Z}_n \quad (18.2.8)$$

$$a + n\mathbb{Z} \mapsto a, \forall a \in \mathbb{Z}_n \quad (18.2.9)$$

Proof. Let us firstly find the distinct cosets of $n\mathbb{Z}$. We have that:

$$a + n\mathbb{Z} = b + n\mathbb{Z} \iff a \in b + n\mathbb{Z} \iff a \equiv b (\text{mod } n) \quad (18.2.10)$$

so the distinct cosets must be:

$$n\mathbb{Z}, 1 + n\mathbb{Z}, \dots, (n - 1) + n\mathbb{Z} \quad (18.2.11)$$

It then follows that ϕ is bijective. Indeed, it is injective since

$$\phi(a + n\mathbb{Z}) = \phi(b + n\mathbb{Z}) \implies a \equiv b (\text{mod } n) \implies a + n\mathbb{Z} = b + n\mathbb{Z} \quad (18.2.12)$$

It is also surjective since:

$$a \in \mathbb{Z}_n \implies a \leq n - 1 \implies \phi(a + n\mathbb{Z}) = a \quad (18.2.13)$$

Finally, for $a, b \in \mathbb{Z}_n$ we find that:

$$\phi((a + n\mathbb{Z}) + (b + n\mathbb{Z})) = \phi((a + b) + n\mathbb{Z}) \quad (18.2.14)$$

$$= \phi(c + n\mathbb{Z}) \quad (18.2.15)$$

$$= c \quad (18.2.16)$$

where $c \equiv a + b \pmod{n}$. Moreover:

$$\phi(a + n\mathbb{Z}) +_n \phi(b + n\mathbb{Z}) = a +_n b = c \quad (18.2.17)$$

so that:

$$\phi((a + n\mathbb{Z}) + (b + n\mathbb{Z})) = \phi(a + n\mathbb{Z}) +_n \phi(b + n\mathbb{Z}) \quad (18.2.18)$$

as desired. ■

Example. Consider the group $\mathbb{Z}/6\mathbb{Z}$. We have established that:

$$\phi : \mathbb{Z}/6\mathbb{Z} \rightarrow \mathbb{Z}_6 \quad (18.2.19)$$

$$a + n\mathbb{Z} \mapsto a, \quad \forall a \in \mathbb{Z}_6 \quad (18.2.20)$$

is an isomorphism. Since \mathbb{Z}_6 is cyclic, it follows that if $\phi(g)$ is a generator of \mathbb{Z}_6 then g must be a generator of $\mathbb{Z}/6\mathbb{Z}$. Now the generators of \mathbb{Z}_6 are 1 and 5 (integers coprime to 6), which are the images of $1 + 6\mathbb{Z}$ and $5 + \mathbb{Z}_6$. The latter two must therefore be generators of $\mathbb{Z}/6\mathbb{Z}$. ◀

18.3 Conjugacy

Definition 19.5 (Conjugacy)

Let $x, y \in G$. Then y is a conjugate of x in G if there exists some $g \in G$ such that:

$$y = gxg^{-1} \quad (18.3.1)$$

Proposition 19.6 (Powers of conjugate elements) Let $x, y, g \in G$ such that $y = gxg^{-1}$. Then $y^n = gx^n g^{-1}$ for all positive integers n .

Proof. We proceed by mathematical induction. Let $P(n) : y^n = gx^n g^{-1}$, then $P(1)$ is clearly true. Moreover, suppose that $P(k)$ is true for some positive integer k . Then:

$$y^k = gx^k g^{-1} \implies y^{k+1} = gx^k g^{-1} y \quad (18.3.2)$$

$$= gx^k g^{-1} gxg^{-1} \quad (18.3.3)$$

$$= gx^k xg^{-1} = gx^{k+1} g^{-1} \quad (18.3.4)$$

so $P(k+1)$ must be true. Hence, by the principle of mathematical induction, we have that $P(n)$ is true for any positive integer n . ■

Theorem 19.7 (Order of conjugate elements)

Let $x, y \in G$ be conjugate elements. Then either x, y have the same finite order or they both have infinite order.

Proof. There exists some $g \in G$ such that $y = gxg^{-1}$. Suppose $x^n = e$, then:

$$y^n = gx^n g^{-1} = geg^{-1} = e \quad (18.3.5)$$

Similarly, suppose that $y^n = e$, then:

$$x^n = g^{-1}y^n g = g^{-1}eg = e \quad (18.3.6)$$

It follows that if there are positive integers n such that $x^n = e$, then $y^n = e$ and vice versa. So either x, y have the same finite order or they have infinite order. ■

Definition 19.8 (Conjugacy class)

Let $x \in G$, then the **conjugacy class** of x in G is the set of all elements in G that are conjugate to x :

$$\{gxg^{-1} : g \in G\} \quad (18.3.7)$$

Theorem 19.9 (Conjugacy class partition)

Let G be a group, then the conjugation relation is an equivalence relation on G . Consequently, the distinct conjugacy classes form a partition of G .

P *R* *e* *f* *l* *e* *x* i *s*: let $x \in G$, then $x = exe^{-1}$, so x is conjugate to itself.

Symmetric: let $x, y \in G$, and suppose x is conjugate to y . That is, there exists some element $g \in G$:

$$x = gyg^{-1} \implies y = g^{-1}xg = g^{-1}x(g^{-1})^{-1} \quad (18.3.8)$$

so it follows that y is conjugate to x .

Transitive: let $x, y, z \in G$, and suppose x is conjugate to y , and y is conjugate to z , so that:

$$x = g_1yg_1^{-1}, \quad y = g_2zg_2^{-1} \quad (18.3.9)$$

for some $g_1, g_2 \in G$. It follows that:

$$x = g_1g_2zg_2^{-1}g_1^{-1} = g_3zg_3^{-1} \quad (18.3.10)$$

where $g_3 = g_1g_2 \in G$. Hence x is conjugate to z .

Since the equivalence classes of an equivalence relation on some set partition the set, we find that the conjugacy classes of G form a partition of the group. ■

It is important to remember that two elements of different order cannot be conjugate to each other. This was proven in Theorem 19.7, and gives us a useful strategy when trying to partition a group into its conjugacy classes. We show this strategy in the example below.

Example. Consider the group $S(\Delta)$. Since the conjugacy classes of this group must contain elements of the same order, we can start by partitioning $S(\Delta)$ into sets of all elements of the same order:

$$\{e\}, \{r, s, t\}, \{a, b\} \quad (18.3.11)$$

where e has order 1, r, s, t have order 2 and a, b have order 3.

Clearly, $\{e\}$ must be a conjugacy class, since there are no other elements of the same order. We can also see that this must be the case by noting that if y is in the conjugacy class of e , $y = geg^{-1} = e$.

Next, let's see if by conjugating r with other elements in $S(\Delta)$ we retrieve s, t . We get that:

$$ara^{-1} = arb = at = s \quad (18.3.12)$$

$$brb^{-1} = bra = bs = t \quad (18.3.13)$$

so we see that $\{r, s, t\}$ is indeed a conjugacy class. Since there are no other elements of order 2 we know that there are no other elements in this class.

Finally, let's see if $\{a, b\}$ is a conjugacy class by the same method. We get that:

$$rar^{-1} = rar = rt = b \quad (18.3.14)$$

so we see that $\{a, b\}$ is indeed another conjugacy class. ◀

For some other groups, such as Abelian groups, there are simpler ways to find the conjugacy partition.

Example. Consider the group $\mathbb{Z}_7^* = \{1, 2, 3, 4, 5, 6\}$.

Since \mathbb{Z}_7^* is an abelian group, it follows that if x is conjugate to y , then $y = x$:

$$g \in G, y = gxg^{-1} = gg^{-1}x = x \quad (18.3.15)$$

Hence, each conjugacy class can only contain one element:

$$\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\} \quad (18.3.16) \quad \blacktriangleleft$$

This example leads us to believe that for any abelian group, the conjugacy classes must

only contain one element.

Proposition 19.10 (Conjugacy classes of abelian groups)

The conjugacy classes of an Abelian group contain only one element each.

Proof. Suppose G is an abelian group, and let $x \in G$. Then, for $g \in G$:

$$gxg^{-1} = gg^{-1}x = ex = x \quad (18.3.17)$$

Therefore, x is only conjugate to itself. Hence the conjugacy class of x is $\{x\}$, as desired. ■

It is important when talking about conjugacy to express what group the elements are conjugate in i.e. the elements x and y are conjugate in the group G . Indeed, suppose $H < G$ is a subgroup, then it is not necessarily true that x, y are conjugate in H . We know that $\exists g \in G$ such that $y = gxg^{-1}$, but we cannot state that $g \in H$. However, the converse is true, that is if x, y are conjugate in H , then x, y must also be conjugate in G .

Proposition 19.11 (Conjugacy in subgroups) Let $H < G$ be a subgroup of some group G , and let $x, y \in H$. Then:

- (i) if x, y are conjugate in H then they are also conjugate in G
- (ii) if x, y are conjugate in G then they are not necessarily conjugate in H

Example. Consider the subgroup $H = \{e, (1 2)(3 4), (1 3)(2 4), (1 4)(2 3)\}$ of S_4 .

Since H has order 4 it must be Abelian, and hence its conjugacy classes can only contain one element. So no two elements of H can be conjugate to each other in H . Yet, because all non-identity elements have the same cyclic structure, they are conjugate to each other in S_4 . ◀

18.4 Normal subgroups and conjugacy

Theorem 19.12 (Normality criteria) Let $H < G$ be a subgroup, then H is normal in G iff:

- (a) $gH = Hg$ for all $g \in G$
- (b) $ghg^{-1} \in H$ for each $h \in H, g \in G$
- (c) $gHg^{-1} = H$ for each $g \in G$
- (d) H is a union of conjugacy classes of G

Proof. The goal of this proof will be to show that the following equivalences and implications hold:

$$(a) \iff (c), (b) \implies (c), (c) \implies (d), (d) \implies (b) \quad (18.4.1)$$

(a) \implies (c) Suppose that $gH = Hg$, $\forall g \in G$. Now suppose $x \in gHg^{-1}$, then $\exists h \in H$ such that:

$$x \in gHg^{-1} \quad (18.4.2)$$

$$\iff x = ghg^{-1} \quad (18.4.3)$$

$$\iff x = h'gg^{-1} = h', h' \in H \implies x \in H \quad (18.4.4)$$

since $gH = Hg$. Therefore we have proven that $gHg^{-1} = H$, as desired.

(c) \implies (a) Suppose that $gHg^{-1} = H$, $\forall g \in G$, and let $g \in G$. Then $\exists h \in H$ such that:

$$x \in gH \quad (18.4.5)$$

$$\iff x = gh \quad (18.4.6)$$

$$\iff x = ghg^{-1}g \quad (18.4.7)$$

$$\iff x = h_1g \implies x \in Hg \quad (18.4.8)$$

Therefore $gH = Hg$ as desired.

(b) \implies (c) Suppose that $ghg^{-1} \in H$, for each $h \in H, g \in G$. Then:

$$h = gg^{-1}hgg^{-1} = g \underbrace{(g^{-1}h(gg^{-1})^{-1})}_{\in H} g^{-1} \quad (18.4.9)$$

Now $(g^{-1}h(gg^{-1})^{-1}) \in H$ by assumption, so that $h \in gHg^{-1}$ as desired. Hence $H \subseteq gHg^{-1}$.

Moreover, since we assumed that $gHg^{-1} \subseteq H$ it follows that $gHg^{-1} = H$ as desired.

(c) \implies (d) Suppose that $gHg^{-1} = H$ for all $g \in G$, and let $h \in H$. Then $ghg^{-1} \in H$ implying that H contains all the conjugates in G of its elements.

(d) \implies (b) Suppose that H is a union of conjugacy classes. Suppose that $h \in H, g \in G$. Then, ghg^{-1} is conjugate to h , and must therefore belong to H . Hence, $ghg^{-1} \in H$.

■

Example. Suppose that H, K are normal subgroups of G . We have proven in Lagrange's theorem that since H, K are subgroups of G , $H \cap K < G$, so let us also prove that $H \cap K \trianglelefteq G$. That is, we need to prove that $ghg^{-1} \in H \cap K$ for all $h \in H \cap K, g \in G$.

Since $H \trianglelefteq G$, we have that $gH = Hg$, $\forall g \in G$, and similarly $gK = Kg$, $\forall g \in G$. Therefore, if we let $x \in H \cap K$ then:

$$gxg^{-1} = ghg^{-1} = h'gg^{-1} = h \quad (18.4.10)$$

for some $h, h' \in H$. Similarly:

$$gxg^{-1} = gkg^{-1} = k'gg^{-1} = k' \quad (18.4.11)$$

for some $k, k' \in K$. It follows then that $gxg^{-1} \in H \cap K$. \blacktriangleleft

Example. Consider the group $X = \{(a, b) \in \mathbb{R}^2 : a \neq 0\}$ equipped with the binary operation:

$$(a, b) * (c, d) = (ac, ad + b) \quad (18.4.12)$$

Consider the subset $K = \{(1, n) : n \in \mathbb{Z}\}$. We firstly prove that this is a subgroup of X .

Closure: suppose $k_1 = (1, n_1), k_2 = (1, n_2) \in K$, then:

$$k_1 * k_2 = (1, n_1) * (1, n_2) = (1, n_1 + n_2) \in K \quad (18.4.13)$$

due to the closure of \mathbb{Z} .

Identity: the identity of X was shown to be $(1, 0)$. This clearly belongs to X , since $0 \in \mathbb{Z}$.

Inverses: the inverse of some element $k = (1, n) \in X$ is $(1, -n)$. This clearly also belongs to K , since $-n \in \mathbb{Z}$ provided $n \in \mathbb{Z}$.

Since the subgroup axioms are satisfied, we have that $K < X$. Now let's see if K is a normal subgroup of X , that is $xkx^{-1} \in K$ for $x \in X, k \in K$. Indeed:

$$xkx^{-1} = (a, b) * (1, n) * \left(\frac{1}{a}, -\frac{b}{a}\right) = (a, b) * \left(\frac{1}{a}, n - \frac{b}{a}\right) \quad (18.4.14)$$

$$= (1, an) \quad (18.4.15)$$

This element does not necessarily belong to K . Indeed, if $a \in \mathbb{R}$ and $n \in \mathbb{Z}$ then an need not to be necessarily an integer. Hence K is not a normal subgroup of X . \blacktriangleleft

Theorem 19.13 (Conjugate subgroup) Let $H < G$ and let $g \in G$. Then $gHg^{-1} < G$.

Proof. Let's check the subgroup axioms.

Closure: let $ghg^{-1}, gkg^{-1} \in gHg^{-1}$, then:

$$(ghg^{-1})(gkg^{-1}) = ghkg^{-1} = gxg^{-1} \quad (18.4.16)$$

where $x = hk \in H$ due to the closure property of subgroups.

Identity: the identity element e in G also belongs to gHg^{-1} , since $e = geg^{-1}$ and $e \in H$.

Inverses: let $ghg^{-1} \in gHg^{-1}$. The inverse of this element in G is $gh^{-1}g^{-1}$, which must also belong to gHg^{-1} since $h^{-1} \in H$.

■

Example. Consider the subgroup $H = \langle s \rangle$ of $S(\square)$. Then, the conjugate subgroup in a is:

$$aHa^{-1} = aHc = a\{e, s\}c = a\{c, t\} = \{e, u\} \quad (18.4.17)$$

◀

We can use the definition of conjugate subgroups for two subgroups. Indeed, if some element g conjugates H to K , then we say that H, K are conjugate subgroups in G .

Proposition 19.14 (Isomorphism of conjugate subgroups)

If H, K are conjugate subgroups in G , then H, K are also isomorphic.

Proof. Suppose H, K are conjugate subgroups in G . Then, $\exists g \in G$ such that $K = gHg^{-1}$. We prove that the following is an isomorphism:

$$\phi : H \rightarrow K \quad (18.4.18)$$

$$h \mapsto ghg^{-1} \quad (18.4.19)$$

This mapping is injective since $\phi(x) = \phi(y)$ implies that $gh_1g^{-1} = gh_2g^{-1} \implies h_1 = h_2$. Moreover, it is surjective since every element of K must be of the form ghg^{-1} where $h \in H$.

Finally:

$$\phi(xy) = gxyg^{-1} = (gxg^{-1})(gyg^{-1}) = \phi(x)\phi(y) \quad (18.4.20)$$

as desired. ■

Example. Consider the following subgroup of A_4 :

$$K = \{e, (1 2)(3 4), (1 3)(2 4), (1 4)(2 3)\} \quad (18.4.21)$$

Let's find the following conjugate subgroups:

$$(1 2 4)K(1 2 4)^{-1} = (1 2 4)K(1 4 2), \text{ and } (2 4 3)K(2 4 3)^{-1} = (2 4 3)K(2 3 4) \quad (18.4.22)$$

Firstly, using the fact that conjugacy does not affect the cycle structure:

$$(1 \ 2 \ 4)e(1 \ 4 \ 2) = (1 \ 2 \ 4)(1 \ 4 \ 2) = e \quad (18.4.23)$$

$$(1 \ 2 \ 4)(1 \ 2)(3 \ 4)(1 \ 4 \ 2) = (1 \ 3)(2 \ 4) \quad (18.4.24)$$

$$(1 \ 2 \ 4)(1 \ 3)(2 \ 4)(1 \ 4 \ 2) = (1 \ 4)(2 \ 3) \quad (18.4.25)$$

$$(1 \ 2 \ 4)(1 \ 4)(2 \ 3)(1 \ 4 \ 2) = (1 \ 2)(3 \ 4) \quad (18.4.26)$$

Similarly:

$$(2 \ 4 \ 3)e(2 \ 3 \ 4) = (2 \ 4 \ 3)(2 \ 3 \ 4) = e \quad (18.4.27)$$

$$(2 \ 4 \ 3)(1 \ 2)(3 \ 4)(2 \ 3 \ 4) = (1 \ 4)(2 \ 3) \quad (18.4.28)$$

$$(2 \ 4 \ 3)(1 \ 3)(2 \ 4)(2 \ 3 \ 4) = (1 \ 2)(3 \ 4) \quad (18.4.29)$$

$$(2 \ 4 \ 3)(1 \ 4)(2 \ 3)(2 \ 3 \ 4) = (1 \ 3)(2 \ 4) \quad (18.4.30)$$

Therefore, $(1 \ 2 \ 4)K(1 \ 2 \ 4)^{-1} = (2 \ 4 \ 3)K(2 \ 4 \ 3)^{-1} = K$.

Since conjugating subgroups must leave the cycle structure invariant, and since there are only three permutations of structure $(- -)(- -)$ in A_4 , it follows that the only conjugating subgroup of K is itself. \blacktriangleleft

Example. Consider the subgroups of $S(\Delta)$:

Order	Subgroups
1	$\{e\}$
2	$\{e, r\}, \{e, s\}, \{e, t\}$
3	$\{e, a, b\}$
6	$S(\Delta)$

and its conjugacy classes which we found earlier:

$$\{e\}, \{a, b\}, \{r, s, t\} \quad (18.4.31)$$

Let's try to find all the normal subgroups of $S(\Delta)$. Recall that $H \trianglelefteq S(\Delta)$ iff H is a union of conjugacy classes of G . We then see that the only subgroups which can be expressed as such unions are:

$$\{e\} = \{e\} \quad (18.4.32)$$

$$\{e, a, b\} = \{e\} \cup \{a, b\} \quad (18.4.33)$$

$$S(\Delta) = \{e\} \cup \{a, b\} \cup \{r, s, t\} \quad (18.4.34)$$

\blacktriangleleft

Unfortunately, often times we do not have a list of all the subgroups of a group. In such cases, it is easier to find which unions of conjugacy classes are normal subgroups. These must contain the conjugacy class $\{e\}$ and they must have order which divides the group's

order, by Lagrange's theorem.

Strategy (*Determining normal subgroups from conjugacy classes*)

- (i) partition G into conjugacy classes
- (ii) Find all the unions of conjugacy classes which include $\{e\}$ and whose order divides $|G|$ as required by Lagrange's theorem.
- (iii) Determine which of these unions are subgroups, and hence normal subgroups.

We illustrate this method below:

Example. These are the conjugacy classes of A_5 :

Conjugacy class	Description	Order
A	$\{e\}$	1
B	(- - -)	20
C	(- -)(- -)	15
D	conjugate to $(1 \ 2 \ 3 \ 4 \ 5)$	12
E	conjugate to $(1 \ 2 \ 3 \ 5 \ 4)$	12

We need to find all possible unions which contain A , whose order divides $|A_5| = 60$, so 1,2,3,4,5,6,10,12,15,20,30,60.

Firstly, the only union which has only one element is $A = \{e\}$.

Secondly, the unions which have 2,3,4,5,6,10 elements do not exist.

Thirdly, the unions which have 12 elements are two, D and E . However this does not contain e , so we scrap it.

Similarly, the only union which has 15 elements is C . However this does not contain e , so we scrap it.

Also, the only union which has 20 elements is B . However this does not contain e , so we scrap it.

There are no unions which contain 30 elements.

Finally, the only union which contains 60 elements is $A \cup B \cup C \cup D \cup E = A_5$.

So we see that the only candidates for normal subgroups are A and A_4 . These are clearly subgroups, and thus also normal subgroups. ◀

18.5 Conjugacy in $S(\mathcal{F})$

In the case of the symmetry groups, we can define conjugacy more concretely by looking at the action of conjugating some symmetry by another symmetry.

Consider for example the effect of $rar^{-1} = c$. We can view this symmetry as applying a on the square that has been reflected about the vertical line of symmetry.

In other words, since we are rotating through $\frac{\pi}{2}$ anti-clockwise on the reflected square, when we reflect back to the original square we see that the overall action of rar^{-1} was to rotate clockwise. Thus, the conjugate symmetry is equivalent to the symmetry we

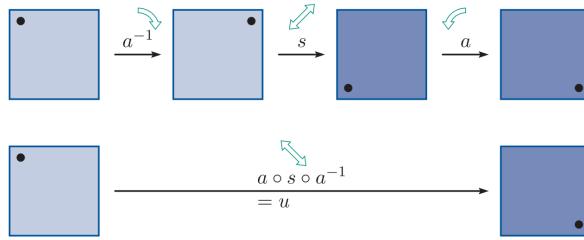


Figure 18.2. Application of rar^{-1}

obtain by applying r to the action of a , that is, reflect the action of a so that it goes from anti-clockwise to clockwise.

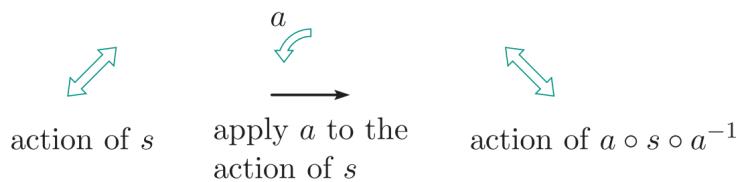


Figure 18.3. Effect of conjugacy on symmetries

More generally, given two symmetries $x, g \in S(\mathcal{F})$, then gxg^{-1} is the symmetry obtained by acting g on the visual effect of x .

It follows that two symmetries are conjugate *iff* there is a symmetry whose action on the visual effect of one of the symmetries is to give the visual effect of the other symmetry.

In other words, if there is a way to relabel the figure in such a way for the effect of the two symmetries to be identical, then they are conjugate.

For example, in the case of r and s , these two symmetries are not conjugate, since there is no possible symmetry of $S(\square)$ that can map the effect of r to the effect of s . We cannot rename the vertices of the square in any possible way for the effect of r and s to coincide.

If instead we considered $S(\text{octagon})$ then we would indeed have that r and s are conjugate symmetries, for example through the rotation by $\frac{\pi}{4}$ anti-clockwise.

Similarly, it is easy to see that r and t are conjugate symmetries. For example, $t = ara^{-1}$, since applying r , a vertical reflection, on a square which has been rotated by $\frac{\pi}{2}$ anti-clockwise, is equivalent to applying t , a horizontal reflection.

Definition 19.15 (Fixed point set)

Let $f \in S(\mathcal{F})$, then the **fixed point set** of f is defined as:

$$\text{Fix } f = \{P \in \mathcal{F} : f(P) = P\} \quad (18.5.1)$$

that is, the subset of \mathcal{F} that is invariant under f .

In two dimensions, for a rotation, the fixed point set consists of the center of rotation. Similarly, for a reflection, the fixed point set consists of the line of reflection.

Since applying g to the visual effect of x gives the action of gxg^{-1} , we expect that if $\text{Fix } f$ are invariant under x , then $g(\text{Fix } f)$ must be the fixed point set of gxg^{-1} .

Theorem 19.16 (Fixed point set of conjugate symmetries)

Let $f, g \in S(\mathcal{F})$, then $\text{Fix } gfg^{-1} = g(\text{Fix } f)$.

Proof. Firstly we prove that $g(\text{Fix } f) \subseteq \text{Fix } gfg^{-1}$. Indeed, suppose that $P \in g(\text{Fix } f)$. Then:

$$g(-1)(P) \in \text{Fix } f \implies (fg^{-1})(P) = g^{-1}(P) \implies (gfg^{-1})(P) = (P) \quad (18.5.2)$$

In other words, $P \in \text{Fix } gfg^1$.

Now, suppose that $P \in \text{Fix } gfg^1$. Then:

$$(gfg^{-1})(P) = P \implies (fg^{-1})(P) = g^{-1}(P) \implies f(g^{-1})(P) = (g^{-1}(P)) \quad (18.5.3)$$

proving that $g^{-1}(P) \in \text{Fix } f$, and thus that $P \in g(\text{Fix } f)$, as desired.

Thus, since $g(\text{Fix } f) \subseteq \text{Fix } gfg^{-1}$ and $\text{Fix } gfg^{-1} \subseteq g(\text{Fix } f)$, we find that $g(\text{Fix } f) = \text{Fix } gfg^{-1}$. ■

This theorem is extremely useful when trying to tell whether or not two symmetries are conjugate. Indeed, the only candidate conjugating symmetries are those that map the fixed point set of one symmetry to the other.

For example, consider the following two symmetries of a tetrahedron:

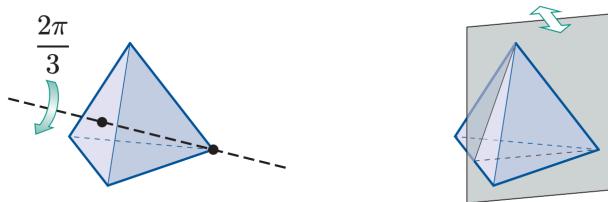


Figure 18.4. Two non-conjugate symmetries of a tetrahedron

The fixed point set of the rotation is simply the axis of rotation, so it is 1-dimensional. The fixed point set of the reflection instead is the plane of reflection, which is 2-dimensional.

It follows immediately that no symmetry can map these two fixed point sets to one another.

More generally, direct symmetries cannot be conjugate to indirect symmetries.

19.17 (Conjugacy direct and indirect symmetries)

A direct symmetry cannot be conjugate to an indirect symmetry.

Proof. Let x be a direct symmetry and y be any symmetry. If g is direct, then gxg^{-1} is also direct. If g is indirect, then gxg^{-1} is direct. Therefore x can only be conjugate to direct symmetries. ■

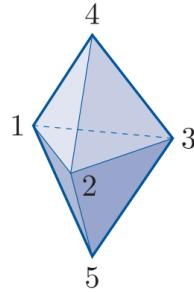
Strategy (Finding conjugacy classes of finite symmetry groups)

- (i) Represent $S(\mathcal{F})$ as a subgroup of a symmetric group.
- (ii) Partition $S(\mathcal{F})$ by cycle structure.
- (iii) For each cycle structure class, determine which of the symmetries are conjugate to each other. R

Recall that the number of elements in each conjugacy class divides $|S(\mathcal{F})|$, and cannot contain both a direct and indirect symmetry.

Also, remember that two symmetries whose fixed point sets cannot be mapped to each other are not conjugate.

Example. Consider the double tetrahedron below:



Firstly, since $|S(\mathcal{F})| = 12$, we see that the conjugacy classes can only contain 1, 2, 3, 4, 6, or 12 elements.

We can see that its symmetries may be categorized in terms of their cycle structure as follows:

$$\{e\} \tag{18.5.4}$$

$$\{(1 2), (1 3), (2 3), (4 5)\} \tag{18.5.5}$$

$$\{(1 2 3), (1 3 2)\} \tag{18.5.6}$$

$$\{(1 2)(4 5), (1 3)(4 5), (2 3)(4 5)\} \tag{18.5.7}$$

$$\{(1 2 3)(4 5), (1 3 2)(4 5)\} \tag{18.5.8}$$

Clearly, $\{e\}$ is a conjugacy class.

Next, we see that $(1\ 2)$, $(1\ 3)$, $(2\ 3)$ are all conjugate through rotations by $\frac{\pi}{3}$ clockwise or anticlockwise, but are not conjugate to $(4\ 5)$.

To check this note that using the renaming method:

$$(2\ 3)(1\ 2)(2\ 3)^{-1} = (1\ 3) \quad (18.5.9)$$

$$(1\ 2)(1\ 3)(1\ 2)^{-1} = (2\ 3) \quad (18.5.10)$$

and since conjugacy is an equivalence relation, transitivity implies that if $(1\ 2)$ is conjugate to $(1\ 3)$, and $(1\ 3)$ is conjugate to $(2\ 3)$, then these must all be conjugate to each other.

To prove that $(4\ 5)$ is not conjugate to the other three, we note that conjugacy does not affect cycle structure. Therefore, if there exists a conjugating symmetry $g \in S(\mathcal{F})$ between $(4\ 5)$ and, say, $(1\ 2)$ then we would need $g = (- -)$ such that:

$$(- -)(4\ 5)(- -) = (1\ 2) \quad (18.5.11)$$

Such a symmetry g does not belong to $S(\mathcal{F})$ (we can check individually all four permutations of structure $(- -)$).

So, we have found two other conjugacy classes, $\{(1\ 2), (1\ 3), (2\ 3)\}$ and $\{(4\ 5)\}$.

Next we see immediately that $(1\ 2\ 3)(4\ 5)$ and $(1\ 3\ 2)(4\ 5)$ are conjugate through $(2\ 3) \in S(\mathcal{F})$:

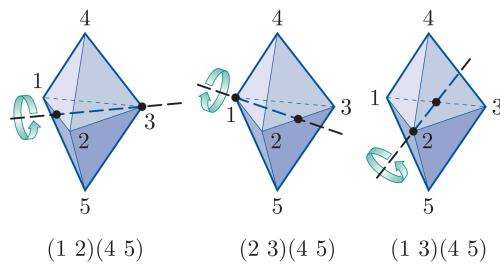
$$(2\ 3)(1\ 2\ 3)(2\ 3)^{-1} = (1\ 3\ 2) \quad (18.5.12)$$

Similarly, we also find that $(1\ 2\ 3)(4\ 5)$ and $(1\ 3\ 2)(4\ 5)$ are conjugate through $(2\ 3) \in S(\mathcal{F})$:

$$(2\ 3)(1\ 2\ 3)(4\ 5)(2\ 3)^{-1} = (1\ 3\ 2)(4\ 5) \quad (18.5.13)$$

Hence we find the conjugacy classes $\{(1\ 2\ 3), (1\ 3\ 2)\}$ and $\{(1\ 2\ 3)(4\ 5), (1\ 3\ 2)(4\ 5)\}$

Finally, $\{(1\ 2)(4\ 5), (1\ 3)(4\ 5), (2\ 3)(4\ 5)\}$ is another conjugacy class. We see this intuitively since these are all rotations about axes in the plane of the triangular base, passing through a vertex and the midpoint of the opposite edge. Therefore, we can conjugate them through rotations by $\frac{\pi}{3}$ clockwise and anti-clockwise.



More rigorously:

$$(2\ 3)(1\ 2)(4\ 5)(2\ 3)^{-1} = (1\ 3)(4\ 5) \quad (18.5.14)$$

$$(1\ 2)(1\ 3)(4\ 5)(1\ 2)^{-1} = (2\ 3)(4\ 5) \quad (18.5.15)$$

as desired. Hence the conjugacy classes of $S(\mathcal{F})$ are:

$$\{e\} \quad (18.5.16)$$

$$\{(1\ 2), (1\ 3), (2\ 3)\} \quad (18.5.17)$$

$$\{(4\ 5)\} \quad (18.5.18)$$

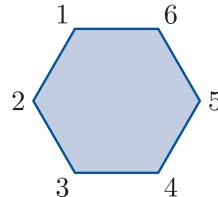
$$\{(1\ 2\ 3), (1\ 3\ 2)\} \quad (18.5.19)$$

$$\{(1\ 2)(4\ 5), (1\ 3)(4\ 5), (2\ 3)(4\ 5)\} \quad (18.5.20)$$

$$\{(1\ 2\ 3)(4\ 5), (1\ 3\ 2)(4\ 5)\} \quad (18.5.21)$$

◀

Example. We label the hexagon as shown:



Consider the symmetries of a hexagon:

$$\{e\} \quad (18.5.22)$$

$$\{(1\ 3)(4\ 6), (2\ 6)(3\ 5), (1\ 5)(2\ 4)\} \quad (18.5.23)$$

$$\{(1\ 4)(2\ 5)(3\ 6), (1\ 6)(2\ 5)(3\ 4), (1\ 2)(3\ 6)(4\ 5), (1\ 4)(2\ 3)(5\ 6)\} \quad (18.5.24)$$

$$\{(1\ 3\ 5)(2\ 4\ 6), (1\ 5\ 3)(2\ 6\ 4)\} \quad (18.5.25)$$

$$\{(1\ 2\ 3\ 4\ 5\ 6), (1\ 6\ 5\ 4\ 3\ 2)\} \quad (18.5.26)$$

Clearly, $\{e\}$ is a conjugacy class.

Instead, $\{(1\ 4)(2\ 5)(3\ 6), (1\ 6)(2\ 5)(3\ 4), (1\ 2)(3\ 6)(4\ 5), (1\ 4)(2\ 3)(5\ 6)\}$ contains one direct symmetry (the first is a rotation by π) and three indirect symmetries which are reflections in axes passing through the midpoints of opposite edges. Hence, we must have the conjugacy class $\{(1\ 4)(2\ 5)(3\ 6)\}$.

Also, $\{(1\ 6)(2\ 5)(3\ 4), (1\ 2)(3\ 6)(4\ 5), (1\ 4)(2\ 3)(5\ 6)\}$ are conjugate through rotations by $\frac{\pi}{3}$ clockwise and anti-clockwise, and the proof is quite similar to the previous example.

Similarly, $\{(1\ 2\ 3\ 4\ 5\ 6), (1\ 6\ 5\ 4\ 3\ 2)\}$ are conjugate through a vertical reflection

$(2\ 6)(3\ 5)$.

Repeating this logical process we quickly see that the conjugacy classes are:

$$\{e\} \quad (18.5.27)$$

$$\{(1\ 4)(2\ 5)(3\ 6)\} \quad (18.5.28)$$

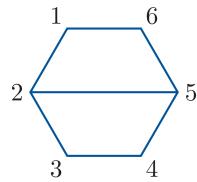
$$\{(1\ 3)(4\ 6), (2\ 6)(3\ 5), (1\ 5)(2\ 4)\} \quad (18.5.29)$$

$$\{(1\ 6)(2\ 5)(3\ 4), (1\ 2)(3\ 6)(4\ 5), (1\ 4)(2\ 3)(5\ 6)\} \quad (18.5.30)$$

$$\{(1\ 3\ 5)(2\ 4\ 6), (1\ 5\ 3)(2\ 6\ 4)\} \quad (18.5.31)$$

$$\{(1\ 2\ 3\ 4\ 5\ 6), (1\ 6\ 5\ 4\ 3\ 2)\} \quad (18.5.32)$$

If we now modify the hexagon as shown below:



The symmetries of the new modified picture form a subgroup of the symmetries of a hexagon. This subgroup can be partitioned into permutations of the same cyclic structure as:

$$\{e\} \quad (18.5.33)$$

$$\{(1\ 3)(4\ 6)\} \quad (18.5.34)$$

$$\{(1\ 6)(2\ 5)(3\ 4)\} \quad (18.5.35)$$



Recall that if G is a group with subgroup H , then H is a normal subgroup of G iff H is a union of conjugacy classes of G . In our case, the symmetry of the modified hexagon definitely does not form a normal subgroup, since it is not the union of conjugacy classes of the normal hexagon's symmetry group.

Unit E3: Homomorphisms

Definition 20.1 (Homomorphism)

Let (G, \circ) and $(H, *)$ be groups. A mapping $\phi : (G, \circ) \rightarrow (H, *)$ is a **homomorphism** if it satisfied the property:

$$\phi(x \circ y) = \phi(x) * \phi(y), \quad \forall x, y \in G \quad (19.0.1)$$

It follows that isomorphisms are homomorphisms that are also bijective.

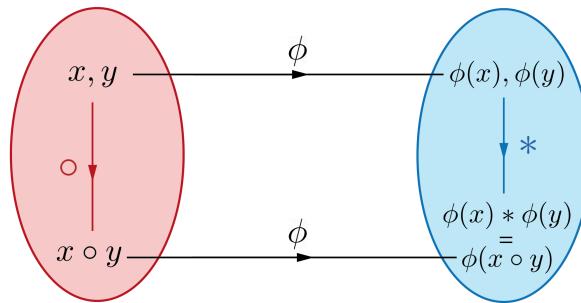


Figure 19.1. Why the homomorphism property must be satisfied for "sensible" maps

Example. Consider the following mapping:

$$\phi : (S_n, \circ) \longrightarrow (\mathbb{Z}_2, +_2) \quad (19.0.2)$$

$$\sigma \mapsto \begin{cases} 0 & \text{if } \sigma \text{ is an even permutation} \\ 1 & \text{if } \sigma \text{ is an odd permutation} \end{cases} \quad (19.0.3)$$

Lets show that it is a homomorphism, thus satisfying $\phi(x \circ y) = \phi(x) +_2 \phi(y), \quad \forall x, y \in S_n$.

Suppose x, y are both of even permutations, then $x \circ y$ is also an even permutation, so that:

$$\phi(x \circ y) = 0 = 0 +_2 0 = \phi(x) +_2 \phi(y) \quad (19.0.4)$$

If instead x, y are both odd permutations, then $x \circ y$ must instead be an even permutation, so that:

$$\phi(x \circ y) = 0 = 1 +_2 1 = \phi(x) +_2 \phi(y) \quad (19.0.5)$$

Finally, if x, y have opposite parity, then $x \circ y$ is an odd permutation, so that:

$$\phi(x \circ y) = 1 = 0 +_2 1 = \phi(x) +_2 \phi(y) \quad (19.0.6)$$

Therefore, ϕ is a homomorphism. \blacktriangleleft

Example. Let us show that the following mapping is not a homomorphism:

$$\phi : (\mathrm{GL}(n, \mathbb{R}), \times) \longrightarrow (\mathrm{GL}(n, \mathbb{R}), \times) \quad (19.0.7)$$

$$A \longmapsto A^{-1} \quad (19.0.8)$$

Consider $A, B \in \mathrm{GL}(n, \mathbb{R})$ then:

$$\phi(A) \times \phi(B) = A^{-1} \times B^{-1} \quad (19.0.9)$$

whereas

$$\phi(A \times B) = (A \times B)^{-1} = B^{-1} \times A^{-1} \quad (19.0.10)$$

However note that $B^{-1} \times A^{-1} \neq A^{-1} \times B^{-1}$ generally. \blacktriangleleft

Proposition 20.2 (\mathbb{Z} homomorphic to \mathbb{Z}_n)

For $n \geq 2$, the following is a homomorphism to:

$$\phi : (\mathbb{Z}, +) \longrightarrow (\mathbb{Z}_n, +_n) \quad (19.0.11)$$

$$k \longmapsto k_{(\text{mod } n)} \quad (19.0.12)$$

where $k_{(\text{mod } n)}$ is the least residue of k modulo n (remained of k when divided by n).

Proof. Suppose $n \geq 2$ and let $r, s \in \mathbb{Z}$. Then:

$$\phi(r + s) = (r + s)_{(\text{mod } n)} \quad (19.0.13)$$

$$\equiv r + s \pmod{n} \quad (19.0.14)$$

$$\equiv r_{(\text{mod } n)} + s_{(\text{mod } n)} \pmod{n} \quad (19.0.15)$$

$$\equiv r_{(\text{mod } n)} +_n s_{(\text{mod } n)} \pmod{n} \quad (19.0.16)$$

$$= \phi(r) +_n \phi(s) \quad (19.0.17)$$

and since $\phi(r + s)$ and $\phi(r) +_n \phi(s)$ both belong to \mathbb{Z}_n , we have that the two must be equal. Thus ϕ is a homomorphism. \blacksquare

Example. Let us prove that:

$$\phi : (G, \circ) \longrightarrow (G, \circ) \quad (19.0.18)$$

$$x \longmapsto x \circ x \quad (19.0.19)$$

is a homomorphism iff (G, \circ) is abelian.

Suppose ϕ is indeed a homomorphism, so that:

$$\phi(x \circ y) = (x \circ y) \circ (x \circ y) = (x \circ x) \circ (y \circ y) = \phi(x) \circ \phi(y) \quad (19.0.20)$$

$$\iff x \circ y \circ x \circ y = x \circ x \circ y \circ y \quad (19.0.21)$$

$$\iff y \circ x = x \circ y \quad (19.0.22)$$

This implies that (G, \circ) is indeed an abelian group, as desired. \blacktriangleleft

Proposition 20.3 (Trivial homomorphism)

Let (G, \circ) and $(H, *)$ be groups, and let e_G, e_H be their respective identity elements. Then the following is a homomorphism:

$$\phi : (G, \circ) \longrightarrow (H, *) \quad (19.0.23)$$

$$x \longmapsto e_H \quad (19.0.24)$$

Proof. Let $x, y \in G$, then:

$$\phi(x \circ y) = e_H = e_H * e_H = \phi(x) * \phi(y) \quad (19.0.25)$$

as desired. \blacksquare

Proposition 20.4 (Properties of homomorphisms)

Let $\phi : (G, \circ) \longrightarrow (H, \odot)$ be a homomorphism, then

$$(i) \text{ let } x_1, x_2, \dots, x_n \in G, \text{ then } \phi\left(\bigodot_{k=1}^n x_k\right) = \bigodot_{k=1}^n \phi(x_k)$$

(ii) $\phi(e_G) = e_H$ where e_G, e_H are the identity elements of G, H respectively.

(iii) for $x \in G$, $\phi(x^{-1}) = (\phi(x))^{-1}$

(iv) for $x \in G$, $\phi(x^n) = (\phi(x))^n$

Proof.

- (i) For the case $n = 1$, it is clear that $\phi(x_1) = \phi(x_1)$. Suppose that for $n \in \mathbb{N}$, we have that:

$$\phi\left(\bigodot_{k=1}^n x_k\right) = \bigodot_{k=1}^n \phi(x_k) \quad (19.0.26)$$

Then:

$$\phi\left(\bigcirc_{k=1}^{n+1} x_k\right) = \phi\left(\left(\bigcirc_{k=1}^n x_k\right) \circ x_{n+1}\right) \quad (19.0.27)$$

$$= \left(\phi\left(\bigcirc_{k=1}^n x_k\right)\right) \odot \phi(x_{n+1}) \quad (19.0.28)$$

$$= \bigcirc_{k=1}^n \phi(x_k) \odot \phi(x_{n+1}) \quad (19.0.29)$$

$$= \bigcirc_{k=1}^{n+1} \phi(x_k) \quad (19.0.30)$$

as desired. Hence by the principle of mathematical induction, we have that $\phi\left(\bigcirc_{k=1}^n x_k\right) = \bigcirc_{k=1}^n \phi(x_k)$.

(ii) We have that $e_G \circ e_G = e_G$, then:

$$\phi(e_G \circ e_G) = \phi(e_G) \quad (19.0.31)$$

giving:

$$\phi(e_G) \odot \phi(e_G) = \phi(e_G) = \phi(e_G) \odot e_H \quad (19.0.32)$$

from which we find that $\phi(e_G) = e_H$.

(ii) Let $x \in G$. Then:

$$x \circ x^{-1} = x^{-1} \circ x = e_G \quad (19.0.33)$$

Then, applying ϕ :

$$\phi(x \circ x^{-1}) = \phi(x^{-1} \circ x) = \phi(e_G) = e_H \quad (19.0.34)$$

Therefore:

$$\phi(x) \odot \phi(x^{-1}) = \phi(x) \odot \phi(x^{-1}) = e_G \quad (19.0.35)$$

implying that $\phi(x^{-1}) = (\phi(x))^{-1}$, as desired.

(iv) We consider two different cases, $n \geq 0$ and $n < 0$.

If $n \geq 0$, then the case $n = 0$ is trivial since:

$$\phi(x^0) = \phi(e_G) = (\phi(x))^0 = e_H \quad (19.0.36)$$

which was proven previously. Now suppose that for some integer $k \geq 0$

$$\phi(x^k) = (\phi(x))^k \quad (19.0.37)$$

Then:

$$\phi(x^{k+1}) = \phi(x^k \circ x) = \phi(x^k) \odot \phi(x) \quad (19.0.38)$$

$$= (\phi(x))^k \odot \phi(x) \quad (19.0.39)$$

$$= (\phi(x))^{k+1} \quad (19.0.40)$$

as desired.

Now suppose $n < 0$, and let $x \in G$. We can write $n = -m$ where $m > 0$. Then:

$$\phi(x^n) = \phi(x^{-m}) \quad (19.0.41)$$

$$= \phi((x^{-1})^m) \quad (19.0.42)$$

$$= (\phi(x^{-1}))^m \quad (19.0.43)$$

$$= (\phi(x))^{-m} \quad (19.0.44)$$

$$= (\phi(x))^n \quad (19.0.45)$$

Therefore, both from cases 1 and 2, we find that:

$$\phi(x^n) = (\phi(x))^n \quad (19.0.46)$$

as desired. ■

Theorem 20.5 (Order of element and homomorphism image)

Let $\phi : (G, \circ) \rightarrow (H, *)$ be a homomorphism and let $x \in G$ be an element of finite order. Then the order of $\phi(x)$ is also finite and divides the order of x .

Proof. We begin by proving the following lemma.

Lemma. Let $x \in G$. If $r > 0$ is a positive integer such that $x^r = e$, then the order of x divides r .

Suppose that $x^r = e$, and suppose that $\text{ord}(x) = s$. Then, we must have that $r = as + b$ for some integers a, b with $0 \leq b < s$.

Hence:

$$e = x^r \quad (19.0.47)$$

$$= x^{as+b} \quad (19.0.48)$$

$$= (x^s)^a \circ x^b \quad (19.0.49)$$

$$= e^a \circ x^b \quad (19.0.50)$$

$$= x^b \quad (19.0.51)$$

Since $b < s$, we find that $b = 0$, or else we would have a contradiction. Therefore, $r = as$, in other words s divides r .

Now since the order of x is s :

$$(\phi(x))^s = \phi(x^s) = \phi(e_G) = e_H \quad (19.0.52)$$

hence the order of ϕ must, by the above lemma, have order that divides s . In other words, the order of $\phi(x)$ must be finite and divide the order of x . ■

Theorem 20.6 (Conjugacy of element and homomorphism image)

Let $\phi : (G, \circ) \rightarrow (H, *)$ be a homomorphism, and let $x, y \in G$. If x, y are conjugate then $\phi(x), \phi(y)$ are conjugate too.

Proof. Suppose x, y are conjugate, so that $\exists g \in G$ such that:

$$y = g \circ x \circ g^{-1} \quad (19.0.53)$$

Therefore:

$$\phi(y) = \phi(g \circ x \circ g^{-1}) = \phi(g) * \phi(x) * \phi(g^{-1}) = \phi(g) * \phi(x) * (\phi(g))^{-1} \quad (19.0.54)$$

so that $\phi(x), \phi(y)$ are also conjugate. ■

19.1 Image and kernels

Definition 20.7 (Image and kernel of homomorphism)

Let $\phi : (G, \circ) \rightarrow (H, *)$ be a homomorphism. Then the **image** of ϕ is:

$$\text{Im } \phi = \{\phi(g) : g \in G\} \quad (19.1.1)$$

and the **kernel** of ϕ is:

$$\ker \phi = \{g \in G : \phi(g) = e_H\} \quad (19.1.2)$$

Theorem 20.8 (Image subgroup of homomorphism)

Let $\phi : (G, \circ) \rightarrow (H, *)$ be a homomorphism. Then $\text{Im } \phi \leq (H, *)$.

Proof. We check the three subgroup axioms:

Closure: let $h_1, h_2 \in \text{Im } \phi$, so that $\exists g_1, g_2 \in G$ satisfying $\phi(g_1) = h_1$ and $\phi(g_2) = h_2$. We find that:

$$h_1 * h_2 = \phi(g_1) * \phi(g_2) = \phi(g_1 \circ g_2) \quad (19.1.3)$$

Therefore, $h_1 * h_2$ is the image of $g_1 \circ g_2$, and thus $h_1 * h_2 \in \text{Im } \phi$.

Identity: the identity of H is e_H , and also belongs to $\text{Im } \phi$ since $\phi(e_G) = e_H$, in other words it is the image of the identity of G .

Inverses: suppose $h \in \text{Im } \phi$. Then, there exists $g \in G$ such that:

$$h = \phi(g) \implies h^{-1} = (\phi(g))^{-1} = \phi(g^{-1}) \quad (19.1.4)$$

so that $h^{-1} \in \text{Im } \phi$.

Hence all three subgroup axioms are satisfied, and thus $\text{Im } \phi \leq (H, *)$. ■

Proposition 20.9 $((G, \circ) \cong \text{Im } \phi)$

Let $\phi : (G, \circ) \rightarrow (H, *)$ be an injective homomorphism and let φ be the map obtained by shrinking the codomain of ϕ to $\text{Im } \phi$. Then φ is an isomorphism, so that $(G, \circ) \cong \text{Im } \phi$.

Proof. φ is still an injective homomorphism, since shrinking the codomain to $\text{Im } \phi$ does not affect the homomorphism property.

However, φ is also surjective, since $\text{Im } \phi = \text{Im } \varphi$ is the domain of this mapping. Consequently, φ is an isomorphism $(G, \circ) \rightarrow \text{Im } \phi$, as desired. ■

Theorem 20.10 (Preserved structures under homomorphism)

Let $\phi : (G, \circ) \rightarrow (H, *)$ be a homomorphism:

- (i) if G is abelian then $(\text{Im } \phi, *)$ is abelian.
- (ii) If G is cyclic then $(\text{Im } \phi, *)$ is cyclic.

Moreover, if G is generated by a , then $(\text{Im } \phi, *)$ is generated by $\phi(a)$.

Proof. (i) Suppose that G is abelian, and let $h_1, h_2 \in \text{Im } \phi$, so that $\exists g_1, g_2$ such that $h_1 = \phi(g_1)$ and $h_2 = \phi(g_2)$. Therefore:

$$\phi(g_1 \circ g_2) = h_1 * h_2 = \phi(g_2 \circ g_1) = h_2 * h_1 \quad (19.1.5)$$

proving that H is also abelian.

(ii) Suppose that $\langle a \rangle = G$, and let $h \in \text{Im } \phi$, so that $h = \phi(g)$ for some $g \in G$. Now since G is generated by a we find that:

$$g = a^k \implies \phi(g) = \phi(a^k) = (\phi(a))^k \quad (19.1.6)$$

Consequently:

$$h = (\phi(a))^k \quad (19.1.7)$$

proving that $\text{Im } \phi$ is generated by $\phi(a)$, and thus also cyclic. ■

Theorem 20.11 (Kernel normal subgroup)

Let $\phi : (G, \circ) \rightarrow (H, *)$ be a homomorphism. Then $\text{Ker}(\phi) \trianglelefteq (G, \circ)$.

Proof. We firstly need to prove that $\text{Ker}(\phi)$ is a subgroup of (G, \circ) by checking the subgroup properties.

Closure: let $k_1, k_2 \in \text{Ker}(\phi)$. Then $\phi(k_1) = e_H$ and $\phi(k_2) = e_H$, so that:

$$\phi(k_1 \circ k_2) = \phi(k_1) * \phi(k_2) = e_H * e_H = e_H \quad (19.1.8)$$

Hence $k_1 \circ k_2 \in \text{Ker}(\phi)$.

Identity: we have $\phi(e_G) = e_H$, so that $e_G \in \text{Ker}(\phi)$.

Inverses: let $k \in \text{Ker}(\phi)$, then $\phi(k) = e_H$ so that:

$$\phi(k^{-1}) = (\phi(k))^{-1} = e_H^{-1} = e_H \quad (19.1.9)$$

so that $k^{-1} \in \text{Ker}(\phi)$.

So, we have that $(\text{Ker}(\phi), \circ) \leq (G, \circ)$.

To prove normality, we must prove that $g \circ k \circ g^{-1} \in \text{Ker}(\phi)$ for $k \in \text{Ker}(\phi)$ and $g \in G$. Indeed:

$$\phi(g \circ k \circ g^{-1}) = \phi(g) * \phi(k) * \phi(g^{-1}) \quad (19.1.10)$$

$$= \phi(g) * e_H * (\phi(g))^{-1} \quad (19.1.11)$$

$$= e_H \quad (19.1.12)$$

as desired, we find that $g \circ k \circ g^{-1} \in \text{Ker}(\phi)$. Hence $(\text{Ker}(\phi), \circ) \trianglelefteq (G, \circ)$ ■

Theorem 20.12 (Injectivity of homomorphism)

Let $\phi : (G, \circ) \rightarrow (H, *)$ be a homomorphism. Then ϕ is injective iff $\text{Ker}(\phi) = \{e_G\}$.

Proof. We begin by proving \Rightarrow . Suppose ϕ is injective, and suppose g_1, g_2 are such that $\phi(g_1) = \phi(g_2) = e_H$. Then, $g_1 = g_2$ due to injectivity, and since $\phi(e_G) = e_H$ it follows that $\text{Ker}(\phi) = \{e_G\}$.

Let us now prove the \Leftarrow part. Suppose that $\text{Ker}(\phi) = \{e_G\}$, and let $\phi(x) = \phi(y)$ for some $x, y \in G$. Then:

$$e_H = \phi(x) * (\phi(y))^{-1} \quad (19.1.13)$$

$$= \phi(x) * \phi(y^{-1}) \quad (19.1.14)$$

$$= \phi(x \circ y^{-1}) \quad (19.1.15)$$

so that $x \circ y^{-1} \in \text{Ker}(\phi) = \{e_G\}$. It follows that $x \circ y^{-1} = e_G \implies x = y$. Hence, ϕ is injective. \blacksquare

Theorem 20.13 (Normality \iff kernel)

Let $K \leq G$, then $K \trianglelefteq G \iff K = \text{Ker}(\phi)$ for some homomorphism ϕ with domain G .

Proof. We begin by proving \implies . Suppose ϕ is a homomorphism with domain G , then by Theorem 20.11 $\text{Ker}(\phi) = K$ is a normal subgroup of G , as desired.

Now let us prove \Leftarrow . Suppose that K is a normal subgroup of G . Then it is the kernel of the map ϕ defined by:

$$\phi : (G, \circ) \longrightarrow (G \setminus K, \cdot) \quad (19.1.16)$$

$$x \mapsto xK \quad (19.1.17)$$

so with domain G , and codomain $G \setminus K$, that is the set of cosets of K in G . Indeed ϕ is a homomorphism since for all $x, y \in G$:

$$\phi(x \circ y) = (x \circ y)K \quad (19.1.18)$$

$$= (xK) \cdot (yK) \quad (19.1.19)$$

$$= \phi(x) \cdot \phi(y) \quad (19.1.20)$$

Also:

$$\text{Ker}(\phi) = \{x \in G : \phi(x) = K\} = \{x \in G : xK = K\} = K \quad (19.1.21) \quad \blacksquare$$

19.2 First isomorphism theorem

Theorem 20.14 (Kernel cosets)

Let $\phi : (G, \circ) \longrightarrow (H, *)$ be a homomorphism, and let $x, y \in G$. Then for some $g \in G$:

$$\phi(x) = \phi(y) \iff x, y \in g\text{Ker}(\phi) \quad (19.2.1)$$

Proof. Firstly, suppose that $\phi(x) = \phi(y)$ for any $x, y \in G$. Then:

$$\phi(x) * \phi(y^{-1}) = \phi(x \circ y^{-1}) = e_H \quad (19.2.2)$$

so that $x \circ y^{-1} \in \text{Ker}(\phi)$. Since $y^{-1} \in G$, it follows that $x \in y\text{Ker}(\phi)$. However, we also have that $y \in y\text{Ker}(\phi)$, since $e_G \in \text{Ker}(\phi)$. Hence x, y both belong to the same coset, with $g = y$. Similarly, we could have also proven that $y \in x\text{Ker}(\phi)$.

Now suppose that x, y lie in the same coset of $\text{Ker}(\phi)$ in G , so that for some $g \in G$ and $k_1, k_2 \in \text{Ker}(\phi)$:

$$x, y \in g\text{Ker}(\phi) \implies x = g \circ k_1, y = g \circ k_2 \implies x = y \circ k_2^{-1} \circ k_1 \quad (19.2.3)$$

Then:

$$\phi(x) = \phi(y \circ k_2^{-1} \circ k_1) = \phi(y) * (\phi(k_2))^{-1} * \phi(k_1) = \phi(y) \quad (19.2.4)$$

as desired. ■

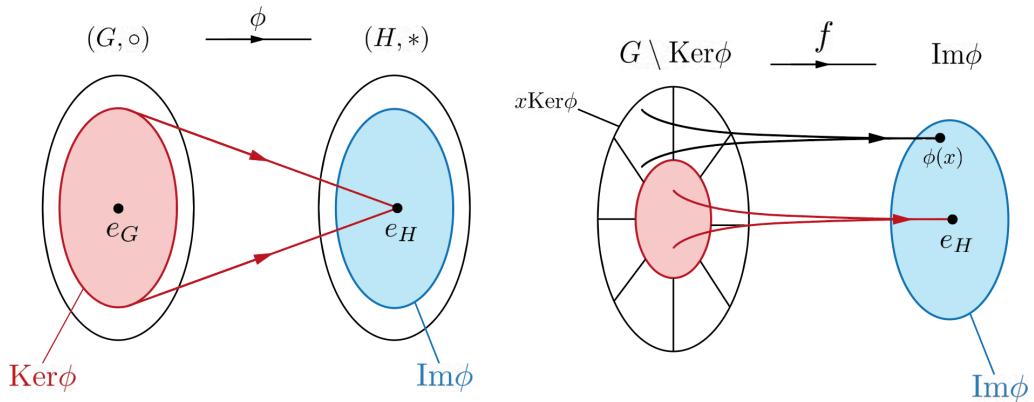


Figure 19.2. Set diagram of first isomorphism theorem

Theorem 20.15 (First isomorphism theorem)

Let $\phi(G, \circ) \rightarrow (H, *)$ be a homomorphism. Then:

$$f : G \setminus \text{Ker}(\phi) \longrightarrow \text{Im}(\phi) \quad (19.2.5)$$

$$x\text{Ker}(\phi) \longmapsto \phi(x) \quad (19.2.6)$$

is an isomorphism, so that $G \setminus \text{Ker}(\phi) \cong \text{Im}(\phi)$.

Proof. For sake of brevity, let $K = \text{Ker}(\phi)$.

Since elements of different cosets of K have different images under ϕ (converse of theorem

20.13), we find that ϕ must be injective. Indeed, suppose that $\phi(x) = \phi(y)$, so that:

$$x \in y\text{Ker}(\phi) = \{yk_1, yk_2, \dots\} \quad (19.2.7)$$

$$y \in x\text{Ker}(\phi) = \{xk_1, xk_2, \dots\} \quad (19.2.8)$$

from which it follows that $x = yk_i$ and $y = xk_i^{-1} = xk_j$ for some $k_i \in \text{Ker}(\phi)$. Therefore, we have that $x\text{Ker}(\phi) \subseteq y\text{Ker}(\phi)$, since if $xk_n \in x\text{Ker}(\phi)$ then $xk_n = yk_i k_n \in y\text{Ker}(\phi)$, using the closure of $\text{Ker}(\phi)$. Similarly, $y\text{Ker}(\phi) \subseteq x\text{Ker}(\phi)$, and thus $x\text{Ker}(\phi) = y\text{Ker}(\phi)$ as desired.

Also, f is surjective, since any element $\phi(x) \in \text{Im}(\phi)$ is the image under f of the coset xK .

Finally, let us check the homomorphism property:

$$f(xK \cdot yK) = f((x \circ y)K) \quad (19.2.9)$$

$$= \phi(x \circ y) \quad (19.2.10)$$

$$= \phi(x) * \phi(y) \quad (19.2.11)$$

$$= f(xK) * f(yK) \quad (19.2.12)$$

as desired. It follows that f is a bijective homomorphism, hence an isomorphism, so that $G \setminus \text{Ker}(\phi) \cong \text{Im}(\phi)$. \blacksquare

Example. Consider the following mapping ϕ :

$$\phi : (L, \times) \longrightarrow (\mathbb{R}^*, \times) \quad (19.2.13)$$

$$\begin{pmatrix} a & 0 \\ b & c \end{pmatrix} \longmapsto ac \quad (19.2.14)$$

where L is the group of lower triangular 2×2 matrices.

This is clearly a homomorphism, since for any $A, B \in L$:

$$A = \begin{pmatrix} a_1 & 0 \\ b_1 & c_1 \end{pmatrix}, \quad B = \begin{pmatrix} a_2 & 0 \\ b_2 & c_2 \end{pmatrix} \quad (19.2.15)$$

then:

$$AB = \begin{pmatrix} a_1 a_2 & 0 \\ a_2 b_1 + c_1 b_2 & c_1 c_2 \end{pmatrix} \quad (19.2.16)$$

Hence:

$$\phi(AB) = \phi \begin{pmatrix} a_1 a_2 & 0 \\ a_2 b_1 + c_1 b_2 & c_1 c_2 \end{pmatrix} \quad (19.2.17)$$

$$= (a_1 a_2)(c_1 c_2) \quad (19.2.18)$$

$$= \phi(A)\phi(B) \quad (19.2.19)$$

as desired.

The image of ϕ is:

$$\text{Im}(\phi) = \{\phi(A) : A \in L\} = \{ac : a, c \in \mathbb{R}^*\} = \mathbb{R}^* \quad (19.2.20)$$

The kernel of ϕ is:

$$\text{Ker}(\phi) = \{A \in L : \det\{A\} = 1\} = \left\{ \begin{pmatrix} a & 0 \\ b & \frac{1}{a} \end{pmatrix} : a \in \mathbb{R}^*, b \in \mathbb{R} \right\} \quad (19.2.21)$$

By the first isomorphism theorem:

$$L \setminus \text{Ker}(\phi) \cong \text{Im}(\phi) = (\mathbb{R}^*, \times) \quad (19.2.22)$$

◀

Example. Consider the following map:

$$\phi : (L, \times) \longrightarrow (L, \times) \quad (19.2.23)$$

$$\begin{pmatrix} a & 0 \\ b & c \end{pmatrix} \longmapsto \begin{pmatrix} \frac{1}{a} & 0 \\ 0 & 1 \end{pmatrix} \quad (19.2.24)$$

This is clearly a homomorphism, since for any $A, B \in L$:

$$A = \begin{pmatrix} a_1 & 0 \\ b_1 & c_1 \end{pmatrix}, \quad B = \begin{pmatrix} a_2 & 0 \\ b_2 & c_2 \end{pmatrix} \quad (19.2.25)$$

then:

$$AB = \begin{pmatrix} a_1 a_2 & 0 \\ a_2 b_1 + c_1 b_2 & c_1 c_2 \end{pmatrix} \quad (19.2.26)$$

Hence:

$$\phi(AB) = \begin{pmatrix} \frac{1}{a_1 a_2} & 0 \\ 0 & 1 \end{pmatrix} \quad (19.2.27)$$

$$= \begin{pmatrix} \frac{1}{a_1} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{a_2} & 0 \\ 0 & 1 \end{pmatrix} \quad (19.2.28)$$

$$= \phi(A)\phi(B) \quad (19.2.29)$$

Moreover it is easy to see that:

$$\text{Im}(\phi) = \left\{ \begin{pmatrix} \frac{1}{a} & 0 \\ 0 & 1 \end{pmatrix} : a \in \mathbb{R}^* \right\} \quad (19.2.30)$$

and:

$$\text{Ker}(\phi) = \left\{ \begin{pmatrix} 1 & 0 \\ b & c \end{pmatrix} : b \in \mathbb{R}, c \in \mathbb{R}^* \right\} \quad (19.2.31)$$

Now, by the first isomorphism theorem, we have that:

$$L \setminus \text{Ker}(\phi) \cong \left\{ \begin{pmatrix} \frac{1}{a} & 0 \\ 0 & 1 \end{pmatrix} : a \in \mathbb{R}^* \right\} \quad (19.2.32)$$

We now prove that $\text{Im}(\phi) \in (\mathbb{R}^*, \times)$. An example of an isomorphism between them is:

$$\varphi : \text{Im}(\phi) \longrightarrow (\mathbb{R}^*, \times) \quad (19.2.33)$$

$$\begin{pmatrix} \frac{1}{a} & 0 \\ 0 & 1 \end{pmatrix} \mapsto a \quad (19.2.34)$$

This is a homomorphism since for $A, B \in \text{Im}(\phi)$:

$$A = \begin{pmatrix} \frac{1}{a} & 0 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} \frac{1}{b} & 0 \\ 0 & 1 \end{pmatrix} \quad (19.2.35)$$

with $a \neq 0, b \neq 0$, we have that

$$\varphi(AB) = ab = \varphi(A)\varphi(B) \quad (19.2.36)$$

Moreover, φ is injective, since $\varphi(A) = \varphi(B) \implies a = b \implies A = B$. Finally, φ is surjective, since every $x \in \mathbb{R}^*$ is the map of the matrix:

$$X = \begin{pmatrix} \frac{1}{x} & 0 \\ 0 & 1 \end{pmatrix} \quad (19.2.37)$$

Hence, we may conclude that $L \setminus \text{Ker}(\phi) \cong (\mathbb{R}^*, \times)$. ◀

Proposition 20.16 (Order of kernel, image and group) Let $\phi : (G, \circ) \longrightarrow (H, *)$ be a homomorphism with a finite group domain, then:

$$|\text{Ker}(\phi)| \cdot |\text{Im}(\phi)| = |G| \quad (19.2.38)$$

Proof. From the first isomorphism theorem, since ϕ is a homomorphism:

$$G \setminus \text{Ker}(\phi) \cong \text{Im}(\phi) \quad (19.2.39)$$

and since isomorphic finite groups must have same order:

$$|G \setminus \text{Ker}(\phi)| = |\text{Im}(\phi)| \quad (19.2.40)$$

Now since G is finite dimensional, we must have that $|G \setminus \text{Ker}(\phi)| = \frac{|G|}{|\text{Ker}(\phi)|}$ so that:

$$|\text{Ker}(\phi)| \cdot |\text{Im}(\phi)| = |G| \quad (19.2.41)$$

as desired. ■

To summarize, we have that for finite groups G, H , and any homomorphism ϕ between them:

- (i) $|\text{Ker}(\phi)|$ divides $|G|$ by Lagrange's theorem, since $\text{Ker}(\phi) \leq G$
- (ii) $|\text{Im}(\phi)|$ divides $|H|$ by Lagrange's theorem, since $\text{Im}(\phi) \leq H$
- (iii) $|\text{Im}(\phi)|$ divides $|G|$ by Proposition 20.16

Example. Let us try to find all homomorphisms ϕ from $S(\Delta)$ to $(\mathbb{Z}_3, +_3)$.

From the above considerations, we must have that $|\text{Ker}(\phi)|$ and $|\text{Im}(\phi)|$ divide $|S(\Delta)| = 6$. Hence they can have values of 1,2,3,6. Moreover, we must have that $|\text{Im}(\phi)|$ divides 3, hence it can only take values 1,3, for which $|\text{Ker}(\phi)|$ takes the values of 6,2 respectively.

Now we know that $|\text{Ker}(\phi)|$ is a normal subgroup of $S(\Delta)$, and we found no normal subgroups of order 2. Hence $|\text{Ker}(\phi)| = 6 = |S(\Delta)|$ and $\text{Im}(\phi) = 1$. Consequently, since $\text{Ker}(\phi) \leq S(\Delta)$ and $\text{Im}(\phi) \leq \mathbb{Z}_3$, we have that $\text{Ker}(\phi) = |S(\Delta)|$ and $\text{Im}(\phi) = \{0\}$. The only possible ϕ with such structures is the trivial homomorphism. ◀

Unit E4: Group actions

20.1 What are group actions?

Several groups that we have considered consist of functions from a set to itself. For example, the elements of the group $S(\square)$ are symmetries, or maps, of the set $\{1, 2, 3, 4\}$ to itself.

We say that when a group element g maps an element x in some set to some other element in the set, then $g : x \mapsto g \wedge x$. In other words, we will denote the image of a set element x under g by $g \wedge x$.

In the case of $S(\square)$, we have for example that $r \wedge 2 = 3$.

For some definitions of \wedge , there are a set of interesting properties which promote it from a simple mapping to a **group action**.

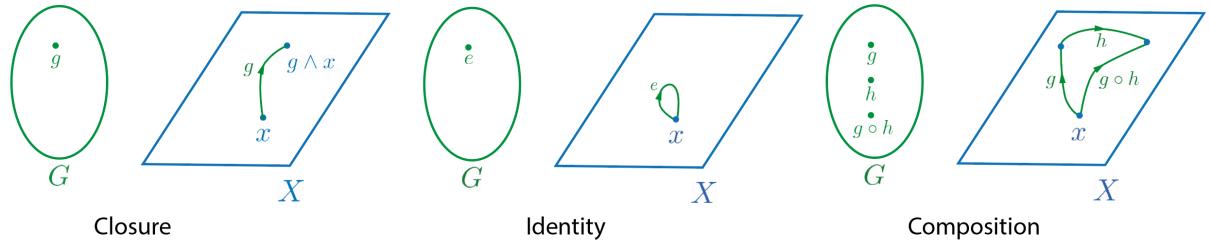


Figure 20.1. Visual illustration of group action axioms.

Definition 21.1 (Group action)

Let (G, \circ) be a group with identity e , and let X be a set. Furthermore, suppose that we associate to each element $g \in G$ and some element $x \in X$ an object $g \wedge x$.

We then say that the effect of \wedge of (G, \circ) on X is a **group action of (G, \circ) on X** , provided that the following properties:

GA1 Closure: for $g \in G$, $x \in X$ we have that $g \wedge x \in X$

GA2 Identity: for $x \in X$, we have that $e \wedge x = x$

GA3 Composition: for $g, h \in G$ and $x \in X$ we have that $g \wedge (h \wedge x) = (g \circ h) \wedge x$. known as the **group action axioms** are satisfied.

Example. Let $G \leq S_5$ be the subgroup consisting of all permutations in S_5 that fix symbols 4 or 5, or transpose them. Consider the set $X = \{1, 2, 3\}$, and let us define:

$$g \wedge x = g(x) \quad (20.1.1)$$

for all $g \in G$, $x \in X$. In other words, G is the set of permutations which does not map 1, 2, 3 to 4 or 5.

We check that the group action axioms are satisfied:

GA1 for $g \in G$, $x \in X$ we have that $g \wedge x = g(x) \in X$, since g must be a permutation of $\{1, 2, 3, 4, 5\}$ which does not map any of $x \in X = \{1, 2, 3\}$ to 4 or 5.

GA2 the identity element of G must be id , the identity permutation, defined so that $\text{id}(x) = \text{id} \wedge x = x$ for all $x \in X$.

GA3 For $g, h \in G$ and $x \in X$ we have that:

$$g \wedge (h \wedge x) = g \wedge h(x) \quad (20.1.2)$$

$$= g(h(x)) = (g \circ h)(x) \quad (20.1.3)$$

$$= (g \circ x) \wedge x \quad (20.1.4)$$

as desired.

It follows that \wedge is indeed a group action of (G, \circ) on X . ◀

Example. Consider now the mapping \wedge of (\mathbb{R}^*, \times) on \mathbb{R}^2 defined by:

$$g \wedge (x, y) = (x + g, y + g) \quad (20.1.5)$$

This is not a group action because it does not satisfy the composition axiom. Indeed $\forall g, h \in (\mathbb{R}^*, \times), \forall (x, y) \in \mathbb{R}^2$ we find that:

$$g \wedge (h \wedge x) = g \wedge (x + h, y + x) = (x + h + g, y + h + g) \quad (20.1.6)$$

whereas:

$$(g \times h) \wedge x = (gh) \wedge x = (x + gh, y + gh) \neq (x + h + g, y + h + g) \quad (20.1.7)$$

as expected. ◀

Theorem 21.2 (Properties of group actions) Let \wedge be an action of G on X , then \wedge is a bijection, that is:

- (i) $\forall g \in G$, if $x, y \in X$ such that $g \wedge x = g \wedge y$ then $x = y$.
- (ii) $\forall g \in G$, if $y \in X$ then $\exists x \in X$ such that $g(x) \in y$

Proof. Let $g \in G$, and suppose we have $x, y \in X$ such that $g \wedge x = g \wedge y$ then $x = y$. Then

we may apply $g^{-1} \in G$:

$$g^{-1} \wedge (g \wedge x) = g^{-1} \wedge (g \wedge y) \implies x = y \quad (20.1.8)$$

using the composition axiom.

Next, let $g \in G$ and $y \in X$. Then:

$$e \wedge y = y \implies g \wedge (g^{-1} \wedge y) = y \quad (20.1.9)$$

and since $g^{-1} \wedge y \in X$ by the closure of group actions, we have that if $x = g^{-1} \wedge y \in X$ then $g \wedge x = y$ as desired. \blacksquare

Theorem 21.3 (Actions of group of symmetries)

Let G be the group of symmetries of some figure $\mathcal{F} \subseteq \mathbb{R}^2$ and let X be a set of figures in \mathbb{R}^2 . Then, if we define \wedge by:

$$g \wedge A = g(A), \forall g \in G, A \in X \quad (20.1.10)$$

then \wedge is a group action *iff* the closure axiom of group actions is satisfied.

Proof. We need to prove that GA2 and GA3 are satisfied.

Firstly let us prove that GA2 holds. Let e be the identity and let $A \in X$. Since $g \in G$ are symmetries of not only \mathcal{F} , but also \mathbb{R}^2 , it follows that e will satisfy $e(P) = P$ for any $P \in \mathbb{R}^2$. Consequently $e \wedge A = e(A) = A$ as desired.

Let $g, h \in G$ and let $A \in X$. Then:

$$g \wedge (h \wedge A) = g(h(A)) = (g \circ h)A \quad (20.1.11)$$

by definition of the composition operation. \blacksquare

Example. Consider now the group $S(\square)$ and the set X whose elements are all the modified 2×2 squares obtained by coloring each of the four small squares either blue, yellow or red.

Then $g \wedge A = g(A)$ for all $A \in X$ is clearly a group action since it satisfies the closure axiom. Indeed, suppose we have a square $A \in X$ with some initial color configuration (c_1, c_2, c_3, c_4) (in clockwise direction starting from top left small square). Then, the action of each element in $S(\square)$ are:

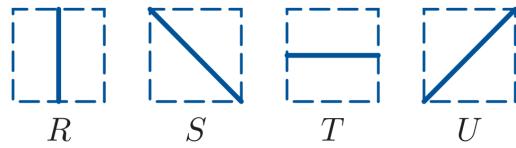
Element	$g \wedge A$
e	(c_1, c_2, c_3, c_4)
a	(c_2, c_3, c_4, c_1)
b	(c_3, c_4, c_1, c_2)
c	(c_4, c_1, c_2, c_3)
r	(c_2, c_1, c_4, c_3)
s	(c_1, c_4, c_3, c_1)
t	(c_4, c_3, c_2, c_1)
u	(c_3, c_2, c_1, c_4)

It is trivial to verify that $g \wedge A \in X$ for all $g \in S(\square)$, thus proving that \wedge is a group action.



An effective way to show the action of a symmetry group on some set is by using cycle notation, as we investigate in the following example.

Example. Consider the action of $S(\square)$ on the set $X = \{R, S, T, U\}$ of figures shown below:



One can easily verify that this is indeed a group action, since the closure property is clearly verified.

We may write down the effect of each element in $S(\square)$ on X using cycle notation:

Element	Permutation
e	i
a	$(R\ T)(S\ U)$
b	i
c	$(R\ T)(S\ U)$
r	$(S\ U)$
s	$(R\ T)$
t	$(S\ U)$
u	$(R\ T)$



20.2 Orbits and stabilisers

Definition 21.4 (Orbit)

Let \wedge be a group action of G on a set X , and let $x \in X$. Then, the **orbit** of x under \wedge is defined as:

$$\text{Orb } x = \{g \wedge x : g \in G\} \subseteq X \quad (20.2.1)$$

that is, the set of elements in X which are the image of x under g .

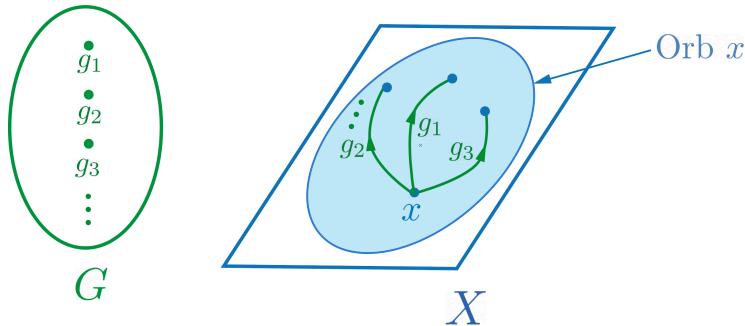


Figure 20.2. Visualization of the orbit of some x under \wedge .

Example. We consider the action of $S(\square)$ on the set $\{1, 2, 3, 4\}$ of labelled vertices of a square. Then:

$$\text{Orb } 1 = \{a \wedge 1, b \wedge 1, c \wedge 1, r \wedge 1, s \wedge 1, t \wedge 1, u \wedge 1\} \quad (20.2.2)$$

$$= \{2, 3, 4, 4, 1, 2, 3\} = \{1, 2, 3, 4\} \quad (20.2.3)$$

Similarly:

$$\text{Orb } 2 = \{a \wedge 2, b \wedge 2, c \wedge 2, r \wedge 2, s \wedge 2, t \wedge 2, u \wedge 2\} \quad (20.2.4)$$

$$= \{3, 4, 1, 3, 4, 1, 2\} = \{1, 2, 3, 4\} \quad (20.2.5)$$

It is easy to verify that $\text{Orb } 3 = \text{Orb } 4 = \{1, 2, 3, 4\}$ also. ◀

Example. Consider the action of $S(\bigcirc)$ on the plane \mathbb{R}^2 , where the disc \bigcirc is placed with its center on the origin.

Then, we find that $\text{Orb } P$ definitely contains the circle C_P centered at the origin passing through P , which is created by acting members of $S^+(\bigcirc)$ on P .

Instead, for the reflections, note that we can obtain all reflection symmetries of the disk by reflecting about the y axis and composing with all direct symmetries. Consequently, we see that the action of the indirect symmetries on P is to create

again a circle centered at the origin passing through P , which is created by acting members of $S^+(\bigcirc)$ on P .

Consequently $\text{Orb } P = \mathcal{C}_P$. ◀

As in the case of conjugacy classes and coset classes, we can prove that orbit classes partition the set X on which the group action acts.

Theorem 21.5 (Orbit partition)

Let \wedge be a group action of (G, \circ) on a set X . Then the distinct orbits of X under \wedge partition X .

Proof. We define \sim on X by:

$$x \sim y \text{ if } y \in \text{Orb } x \quad (20.2.6)$$

and prove that it is an equivalence relation.

Indeed:

- (i) **Reflexivity:** let $x \in X$, then we see that $x \in \text{Orb } x$ since $x = e \wedge x$. Hence $x \sim x$.
- (ii) **Symmetry:** let $x, y \in X$ such that $x \sim y$. Then $y \in \text{Orb } x$, that is $y = g \wedge x$ for some $g \in G$. Then:

$$g^{-1} \wedge y = (g^{-1} \circ g) \wedge x = x \implies x \in \text{Orb } y \quad (20.2.7)$$

since $g^{-1} \in G$. Consequently $y \sim x$.

- (iii) **Transitivity:** let $x, y, z \in X$ such that $x \sim y$ and $y \sim z$, or equivalently $y \in \text{Orb } x$ and $z \in \text{Orb } y$. We can write these as:

$$y = g_1 \wedge x, \text{ and } z = g_2 \wedge y \quad (20.2.8)$$

for some $g_1, g_2 \in G$. Then:

$$z = g_2 \wedge (g_1 \wedge x) = (g_2 \circ g_1) \wedge x \implies z \in \text{Orb } x \quad (20.2.9)$$

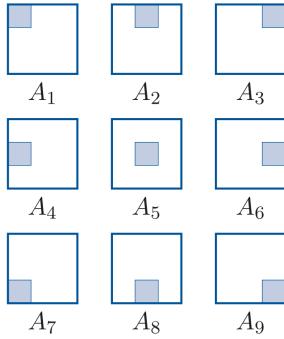
since $g_2 \circ g_1 \in G$. Consequently $x \sim z$ as desired.

It follows that \sim is an equivalence relation, and that its equivalence classes thus partition X . The equivalence classes may be expressed as:

$$[x] = \{y : y \in \text{Orb } x\} = \text{Orb } x \quad (20.2.10)$$

so it follows that the distinct orbits of an element $x \in X$ under \wedge partition X . ■

Example. Consider the action of $S(\square)$ on the set $X = \{A_i : 1 \leq i \leq 9\}$ of squares shown below:



We begin by writing down $\text{Orb } A_1$:

$$\text{Orb } A_1 = \{A_1, A_7, A_9, A_3\} \quad (20.2.11)$$

Next, choosing an element that was already included would have resulted in the same set. Hence, we see that A_2 was not included in the above orbit, so we find its orbit, :

$$\text{Orb } A_2 = \{A_2, A_4, A_8, A_6\} \quad (20.2.12)$$

The only remaining element of X is A_5 , so its orbit must only contain itself:

$$\text{Orb } A_5 = \{A_5\} \quad (20.2.13)$$

Hence the orbit partition of X is:

$$X = \{A_1, A_3, A_7, A_9\} \cup \{A_2, A_4, A_6, A_8\} \cup \{A_5\} \quad (20.2.14)$$

◀

Example. Consider the matrix group

$$G = \left\{ \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} : a, b \in \mathbb{R}^+ \right\} \quad (20.2.15)$$

and the group action \wedge on \mathbb{R}^2 defined by:

$$\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \wedge (x, y) = (ax, by) \quad (20.2.16)$$

for all $(x, y) \in \mathbb{R}^2$. It follows that:

$$\text{Orb } (x, y) = \{(ax, by) : a, b \in \mathbb{R}^+\} \quad (20.2.17)$$

For example:

$$\text{Orb } (1, 0) = \{(a, 0) : a \in \mathbb{R}^+\} \quad (20.2.18)$$

which is the positive x -axis excluding the origin. By symmetry, $\text{Orb } (-1, 0)$ must be the negative x -axis excluding the origin.

Similarly:

$$\text{Orb } (0, -1) = \{(0, -b) : b \in \mathbb{R}^+\} \quad (20.2.19)$$

which is the negative y -axis excluding the origin. By symmetry, $\text{Orb } (0, 1)$ must be the positive y -axis excluding the origin.

One element of \mathbb{R}^2 which we did not include is the origin. We can guess that it is the orbit of the origin, indeed:

$$\text{Orb } (0, 0) = \{(0, 0)\} \quad (20.2.20)$$

We are missing the four quadrants of \mathbb{R}^2 . Note that

$$\text{Orb } (1, 1) = \{(a, b) \in \mathbb{R}^+ : a, b \in \mathbb{R}\} \quad (20.2.21)$$

which is the upper right quadrant excluding the origin and the axes. By symmetry, $\text{Orb } (1, -1)$ must be the lower right quadrant, $\text{Orb } (-1, 1)$ must be the upper left quadrant and $\text{Orb } (-1, -1)$ must be the lower left quadrant. \blacktriangleleft

Definition 21.6 (Stabiliser)

Let \wedge be an action of a group G on a set X , and let $x \in X$. Then, the **stabiliser** of x under \wedge is defined as:

$$\text{Stab } x = \{g \in G : g \wedge x = x\} \quad (20.2.22)$$

We may interpret the stabiliser geometrically as shown below:

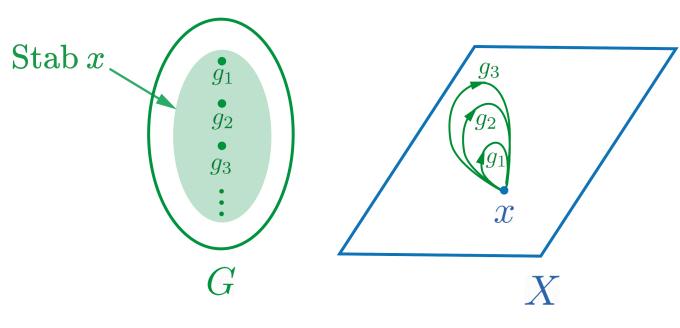


Figure 20.3. Geometrical interpretation of the stabiliser

Example. Let's consider the action of $S(\bigcirc)$ on \mathbb{R}^2 , where the disc \bigcirc is centered at the origin, and let $P \in \mathbb{R}^2$. We consider the following possibilities:

- (i) P is the origin: then $\text{Stab } P = S(\bigcirc)$
- (ii) P is not the origin: then $\text{Stab } P = \{e, q\}$ contains e and the reflection in the line containing P and O , which we call q .

◀

Interestingly, the stabilisers we have found in the previous example can be verified to be subgroups of $S(\bigcirc)$. This is no coincidence, as the following theorem shows.

Theorem 21.7 (Stabiliser subgroup)

Let \wedge be an action of (G, \circ) on X . Then, for any $x \in X$, $\text{Stab } x \leq G$.

Proof. We show that the subgroup axioms are satisfied:

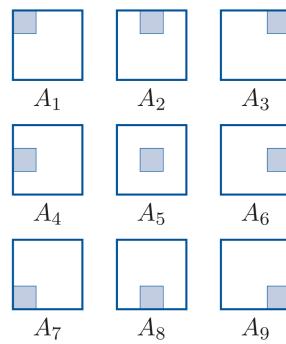
Closure: let $g, h \in \text{Stab } x$. Then $(g \circ h) \wedge x = g \wedge (h \wedge x) = g \wedge x = x$ so $g \circ h \in \text{Stab } x$ as desired.

Identity: let e be the identity of G . Then $e \wedge x = x$ by the group action axioms, hence $e \in \text{Stab } x$.

Inverses: let $g \in \text{Stab } x$, then $g \wedge x = x \implies x = g^{-1} \wedge x$ so that $g^{-1} \in \text{Stab } x$.

■

Example. Consider the action of $S(\square)$ on the set $X = \{A_i : 1 \leq i \leq 9\}$ of squares shown below:



Then:

$$\text{Stab } A_1 = \{e, s\} = \text{Stab } A_9 \quad (20.2.23)$$

$$\text{Stab } A_2 = \{e, r\} = \text{Stab } A_8 \quad (20.2.24)$$

$$\text{Stab } A_3 = \{e, u\} = \text{Stab } A_7 \quad (20.2.25)$$

$$\text{Stab } A_4 = \{e, t\} = \text{Stab } A_6 \quad (20.2.26)$$

$$\text{Stab } A_5 = S(\square) \quad (20.2.27)$$

These are all subgroups of $S(\square)$, since the first four lines contain e and a reflection (which are self inverse). \blacktriangleleft

Example. Consider the action of $G = \left\{ \begin{pmatrix} a & b \\ 0 & a \end{pmatrix} : a, b \in \mathbb{R}, a \neq 0 \right\}$ on \mathbb{R}^2 defined by:

$$\begin{pmatrix} a & b \\ 0 & a \end{pmatrix} \wedge (x, y) = (ax, ay), \quad \forall (x, y) \in \mathbb{R}^2 \quad (20.2.28)$$

Then, we have that if (x, y) is not the origin:

$$\text{Stab } (x, y) = \left\{ \begin{pmatrix} a & b \\ 0 & a \end{pmatrix} \in G : (ax, ay) = (x, y) \right\} = \left\{ \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} : b \in \mathbb{R} \right\} \quad (20.2.29)$$

If instead $(x, y) = (0, 0)$ then:

$$\text{Stab } (0, 0) = \left\{ \begin{pmatrix} a & b \\ 0 & a \end{pmatrix} \in G : (a \cdot 0, a \cdot 0) = (0, 0) \right\} = G \quad (20.2.30)$$

\blacktriangleleft

20.3 The Orbit-Stabiliser theorem

Theorem 21.8 (Left coset of stabiliser) Let \wedge be an action of (G, \circ) on a set X , and let $x \in X$, and $g, h \in G$, then:

$$g \wedge x = h \wedge x \iff g, h \text{ lie in the same left coset of Stab } x \quad (20.3.1)$$

Proof. We firstly prove \implies . Indeed, suppose g, h lie in the same left coset of Stab x , so that $h \in \text{Stab } x$. Then $\exists k \in \text{Stab } x$ such that $h = g \circ k$ and thus:

$$h \wedge x = g \wedge k \wedge x = g \wedge x \quad (20.3.2)$$

as desired.

Now suppose that $h \wedge x = g \wedge x$. Then:

$$(g^{-1} \circ h) \wedge x = g^{-1} \wedge (h \wedge x) \quad (20.3.3)$$

$$= g^{-1} \wedge (g \wedge x) \quad (20.3.4)$$

$$= (g^{-1} \circ g) \wedge x \quad (20.3.5)$$

$$= e \wedge x \quad (20.3.6)$$

$$= x \quad (20.3.7)$$

so that $g^{-1} \circ h \in \text{Stab } x$, and consequently $h \in g\text{Stab } x$, so that h, g both lie in the same left coset. \blacksquare

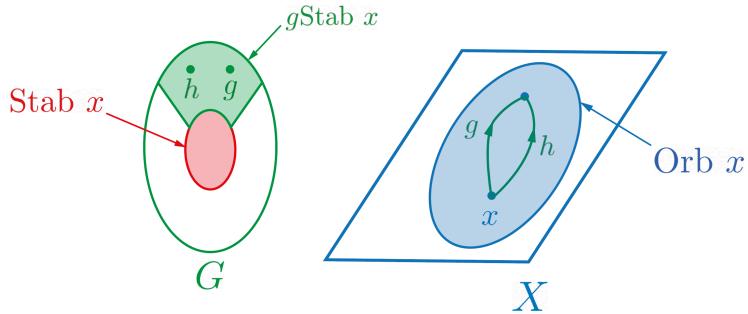


Figure 20.4. Visual interpretation of the left coset $g\text{Stab } x$ and its action on X

Example. We consider the action of S_3 on $\{1, 2, 3\}$, and find that:

$$\text{Stab } 1 = \{\sigma \in S_3 : \sigma(1) = 1\} = \{e, (2 \ 3)\} \quad (20.3.8)$$

Then, the left cosets of this stabiliser are:

$$e\text{Stab } 1 = \text{Stab } 1 \quad (20.3.9)$$

$$(1 \ 2)\text{Stab } 1 = \{(1 \ 2), (1 \ 2 \ 3)\} \quad (20.3.10)$$

$$(1 \ 3)\text{Stab } 1 = \{(1 \ 3), (1 \ 3 \ 2)\} \quad (20.3.11)$$

$$(2 \ 3)\text{Stab } 1 = \text{Stab } 1 \quad (20.3.12)$$

$$(1 \ 2 \ 3)\text{Stab } 1 = \{(1 \ 2), (1 \ 2 \ 3)\} \quad (20.3.13)$$

$$(1 \ 3 \ 2)\text{Stab } 1 = \{(1 \ 3), (1 \ 3 \ 2)\} \quad (20.3.14)$$

Now we partition S_3 according to where its elements map 1:

$$\{e, (2 \ 3)\} \text{ map 1 to 1} \quad (20.3.15)$$

$$\{(1 \ 2), (1 \ 2 \ 3)\} \text{ map 1 to 2} \quad (20.3.16)$$

$$\{(1 \ 3), (1 \ 3 \ 2)\} \text{ map 1 to 3} \quad (20.3.17)$$

which are precisely the left cosets we calculated earlier. ◀

Proposition 21.9 (Stabiliser coset to orbit map)

Let \wedge be an action of G on the set X and let $x \in X$. Then:

$$\phi : G\text{Stab } x \longrightarrow \text{Orb } x \quad (20.3.18)$$

$$g\text{Stab } x \mapsto g \wedge x \quad (20.3.19)$$

is a bijective map.

Proof. Since elements of different cosets of $\text{Stab } x$ have different images under ϕ , we find that ϕ must be injective. Indeed, suppose that $\phi(g\text{Stab } x) = \phi(h\text{Stab } x)$, so that $g \wedge x = h \wedge x$:

$$g \in h\text{Stab } x = \{hs_1, hs_2, \dots\} \quad (20.3.20)$$

$$h \in g\text{Stab } x = \{gs_1, gs_2, \dots\} \quad (20.3.21)$$

from which it follows that $g = hs_i$ and $h = gs_i^{-1} = gs_j$ for some $s_i, s_j \in \text{Stab } x$. Therefore, we have that $g\text{Stab } x \subseteq h\text{Stab } x$, since if $gs_n \in g\text{Stab } x$ then $gs_n = hs_is_n \in h\text{Stab } x$, using the closure of $\text{Stab } x$. Similarly, $h\text{Stab } x \subseteq g\text{Stab } x$, and thus $g\text{Stab } x = h\text{Stab } x$ as desired.

Also, f is surjective, since any element $g \wedge x \in \text{Im}(\phi)$ is the image under f of the left coset $x\text{Stab } x$. ■

Theorem 21.10 (Orbit-Stabiliser theorem)

Suppose that G is a finite group acting on the set X . Then:

$$\forall x \in X, |\text{Orb } x| \times |\text{Stab } x| = |G| \quad (20.3.22)$$

Proof. Let $x \in X$, we know from Proposition 21.9 that the left cosets of $\text{Stab } x$ in G have a bijective correspondence with the elements of $\text{Orb } x$. It follows that $|G\text{Stab } x| = |\text{Orb } x|$, the number of distinct left cosets of $\text{Stab } x$ is equal to the number of elements in $\text{Orb } x$. However, since G is a finite group $|G\text{Stab } x| = \frac{|G|}{|\text{Stab } x|}$ so that:

$$|\text{Orb } x| \cdot |\text{Stab } x| = |G| \quad (20.3.23)$$

■

Interestingly, our choice of X is not limited to sets. Indeed, we can consider group actions on groups themselves, in other words X can be a group.

One example of such a group action is conjugation.

Proposition 21.11 (Conjugation group action)

Let G be a group with $g, x \in G$ and define \wedge by:

$$g \wedge x = gxg^{-1} \quad (20.3.24)$$

Then \wedge is a group action.

Proof. We prove that the three group action axioms hold.

GA1 Closure: let $g, x \in G$. Then $g \wedge x = gxg^{-1} \in G$

GA2 Identity: let $x \in G$ and let $e \in G$ be the identity element. Then $e \wedge x = exe^{-1} = x$ as desired.

GA3 Composition: let $g, h, x \in G$. Then:

$$g \wedge (h \wedge x) = g \wedge (hxh^{-1}) \quad (20.3.25)$$

$$= g(hxh^{-1})g^{-1} \quad (20.3.26)$$

$$= (gh)x(gh)^{-1} \quad (20.3.27)$$

$$= (gh) \wedge x \quad (20.3.28)$$

as desired. ■

Example. We prove that $h \wedge g = hg$ is a group action, where $h \in H, g \in G$ and $H \leq G$.

Indeed:

GA1 Closure: let $g \in G$ and $h \in H \implies h \in G$. Then $g \wedge h = gh \in G$ as desired.

GA2 Identity: let $g \in G$ and let $e \in H$ be the identity element of H , and thus of G too. Then $e \wedge g = eg = g$ as desired.

GA3 Composition: let $g, h, f \in G$. Then:

$$f \wedge (h \wedge g) = f \wedge (hx) \quad (20.3.29)$$

$$= fhx \quad (20.3.30)$$

$$= (fh)x \quad (20.3.31)$$

$$= (fh) \wedge x \quad (20.3.32)$$

where we used the associativity in G . ◀

Proposition 21.12 (Cardinality of conjugacy class)

For a finite group G , the number of elements in each conjugacy class divides $|G|$.

Proof. Let G be a finite group and let \wedge be the conjugacy action $g \wedge x = gxg^{-1}$ for $g, x \in G$. Then:

$$\text{Orb } x = \{gxg^{-1} : g \in G\} = [x] \quad (20.3.33)$$

so the orbit of x is the conjugacy class of x . We also have that:

$$\text{Orb } x \text{ divides } |G| \quad (20.3.34)$$

giving the desired result. ■

Proposition 21.13 (Homomorphism group action)

Let $\phi : (G, \circ) \rightarrow (H, *)$ be a homomorphism, and let \wedge be defined as:

$$g \wedge h = \phi(g) * h \quad (20.3.35)$$

for $g \in G, h \in H$. Then we have that \wedge is a group action.

Proof. We show that the three group action axioms are satisfied:

GA1 Closure: let $g \in G, h \in H$, then $g \wedge h = \phi(g) * h \in H$ since $\phi(g) \in H$.

GA2 Identity: let $e_G \in G$ be the identity of G and let $h \in H$. Then $e_G \wedge h = \phi(e_G) * h = e_H * h = h$ as desired.

GA3 Composition: let $g, f \in G$ and $h \in H$. Then:

$$g \wedge (f \wedge h) = g \wedge (\phi(f) * h) \quad (20.3.36)$$

$$= \phi(g) * (\phi(f) * h) \quad (20.3.37)$$

$$= \phi(g \circ f) * h \quad (20.3.38)$$

$$= (g \circ f) \wedge h \quad (20.3.39)$$

as desired. ■

Notice that for the homomorphism group action of G on H :

$$\text{Orb } e_H = \{\phi(g) * e_H : g \in G\} = \{\phi(g) : g \in G\} = \text{Im}(\phi) \quad (20.3.40)$$

and:

$$\text{Stab } e_H = \{g : \phi(g) * e_H = e_H\} = \{g : \phi(g) = e_H\} = \text{Ker}(\phi) \quad (20.3.41)$$

Applying the orbit stabiliser theorem:

$$|\text{Orb } e_H| \cdot |\text{Stab } e_H| = |\text{Im}(\phi)| \cdot |\text{Ker}(\phi)| = |G| \quad (20.3.42)$$

which is precisely the result proven in Proposition 20.16.

20.4 The Counting theorem

Consider a 2×2 square pattern, where each of the four smaller squares is colored either blue, yellow, red, green or purple?

We see that since repetitions are allowed, each tile has 5 different possible colours. Therefore, we should have $5^4 = 625$ different patterns.

However, note that the following two patterns, which are rotations of each other, were counted twice in our procedure:

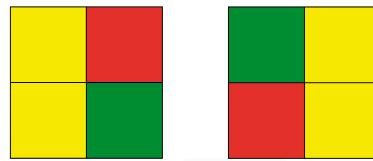


Figure 20.5. Two identical patterns which were double counted.

Surprisingly, this problem has to do with group actions. Indeed, let X be the set of 5^4 colored squares, where each small square is fixed in space. We can think of these squares as the vertices of a larger square, with symmetry group $S(\square)$. Two patterns which are double counted are therefore in the same orbit of the action of $S(\square)$ on X .

We can reformulate this square coloring problem as finding the number of orbits of the action of $S(\square)$ on X .

Definition 20.17 (Fixed set)

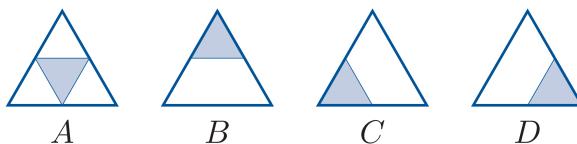
Let \wedge be a group action of G on X , and let $g \in G$. Then the **fixed set** of g under \wedge is defined as:

$$\text{Fix } g = \{x \in X : g \wedge x = x\} \quad (20.4.1)$$

that is, the subset of X whose elements are mapped to themselves by g .

Example. Consider the action of $S(\Delta)$ on the set $X = \{A, B, C, D\}$ of triangles

shown below:



Then, we see that:

$$\text{Fix } e = X \quad (20.4.2)$$

$$\text{Fix } a = \text{Fix } b = \{A\} \quad (20.4.3)$$

$$\text{Fix } r = \{A, B\} \quad (20.4.4)$$

$$\text{Fix } s = \{A, C\} \quad (20.4.5)$$

$$\text{Fix } t = \{A, D\} \quad (20.4.6)$$

◀

Example. Let

$$G = \left\{ \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} : a, b \in \mathbb{R}, a \neq 0 \right\} \quad (20.4.7)$$

and consider the action of G on \mathbb{R}^2 defined by:

$$\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} \wedge (x, y) = (ax, y) \quad (20.4.8)$$

Then, since $a \neq 0$ it follows that

$$\text{Fix } \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} = \{(x, y) \in \mathbb{R}^2 : (x, y) = (ax, y)\} = \{(0, y) : y \in \mathbb{R}\} \quad (20.4.9)$$

if $a \neq 1$ and

$$\text{Fix } \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} = \mathbb{R}^2 \quad (20.4.10)$$

if $a = 1$.

◀

Let's use the notion of fixed sets in the context of the 2×2 square pattern problem.

Example. Let's find the number of elements in the fixed sets of each symmetry in $S(\square)$.

Clearly, $\text{Fix } e = X$, so that $|\text{Fix } e| = 5^4$.

Moreover, $\text{Fix } a$ is the set of squares where all squares are colored in the same way. To see why this must be the case, let us number the four squares in a pattern in $\text{Fix } a$ by c_1, c_2, c_3, c_4 representing their colors. If we rotate this pattern, we find that:

$$\begin{array}{|c|c|} \hline c_1 & c_2 \\ \hline c_4 & c_3 \\ \hline \end{array} \xrightarrow{a} \begin{array}{|c|c|} \hline c_2 & c_3 \\ \hline c_1 & c_4 \\ \hline \end{array}$$

so that $c_1 = c_2, c_4 = c_1, c_3 = c_4, c_2 = c_3$ implying that all colors must be the same.

There are 5 such squares.

By similar arguments, we can see that $\text{Fix } b$ must be the set of squares where the diagonal small squares are of the same color. Note that the color of small squares on different diagonals need not to be the same, hence there are 5^2 such squares in this fixed set.

Similarly, $\text{Fix } c$ contains 5 elements, $|\text{Fix } r| = |\text{Fix } t| = 5^2$. Finally, $|\text{Fix } s| = |\text{Fix } u| = 5^3$. We summarize these results in the table below:

$g \in S(\square)$	$ \text{Fix } g $
e	5^4
a	5
b	5^2
c	5
r	5^2
s	5^3
t	5^2
u	5^3



Theorem 20.18 (Counting theorem)

Let \wedge be an action of a finite group G on the set X , then the number of orbits of \wedge is:

$$\frac{1}{|G|} \sum_{g \in G} |\text{Fix } g| \quad (20.4.11)$$

Proof. Let t be the number of orbits, and let B be one such orbit. Then:

$$\sum_{x \in B} |\text{Stab } x| = \sum_{x \in B} \frac{|G|}{|\text{Orb } x|} \quad (20.4.12)$$

$$= |G| \sum_{x \in B} \frac{1}{|\text{Orb } x|} \quad (20.4.13)$$

$$= |G| \sum_{x \in B} \frac{1}{|B|} \quad (20.4.14)$$

$$= |G| \cdot |B| \cdot \frac{1}{|B|} \quad (20.4.15)$$

$$= |G| \quad (20.4.16)$$

Therefore, since orbits partition X :

$$\sum_{x \in X} |\text{Stab } x| = t|G| \iff t = \frac{1}{|G|} \sum_{x \in X} |\text{Stab } x| \quad (20.4.17)$$

However, we have that: $\sum_{x \in X} |\text{Stab } x| = \sum_{g \in G} |\text{Fix } g|$

Indeed, suppose we construct a table with a row heading containing $x \in X$ and with a column heading containing $g \in G$. We place a y at each position where $g \wedge x = x$. Then $\sum_{x \in X} |\text{Stab } x|$ corresponds in counting the number of ticks in each column labelled x , and summing them all up. This surely must be equivalent to counting the number of ticks in each row labelled g and summing them all up, which corresponds to $\sum_{g \in G} |\text{Fix } g|$.

	... x ...
⋮	⋮
g	... ✓ ✓ ... ✓ ... ✓ ...
⋮	⋮
	✓
	✓
	⋮

■

Example. Let us return to the problem of coloring a 2×2 square. We need to be careful in defining what we mean by two squares being the same. In this particular example, we consider two squares as being the same if one can be rotated or flipped to give the other. We therefore need to find the number of orbits of the action of $S(\square)$:

$$\frac{1}{8}(5^4 + 5 + 5^2 + 5 + 5^2 + 5^3 + 5^2 + 5^3) = 120 \quad (20.4.18)$$

so there are 120 different patterns. ◀

Let us apply our results to one final coloring problem.

Example. Let's see how many different ways there are to color a cube's faces using three colors. We consider two cubes identical if one can be rotated to give the other (but not reflected, obviously).

First, we need to find the fixed sets of each element in $S^+(\text{cube})$ (we do not consider reflections). The elements in $S^+(\text{cube})$ are:

- (a) identity symmetry
- (b) rotations by $\pm\frac{\pi}{2}$ about axes through centers of opposite faces
- (c) rotations by π about axes through centers of opposite faces
- (d) rotations by $\pm\frac{2\pi}{3}$ about axes through opposite vertices
- (e) rotations by π about axes through midpoints of opposite edges.

We see that there is 1 symmetry of type a, 6 rotations of type b, 3 rotations of type c, 8 rotations of type d and 6 rotations of type e.

For each, we find through the labelling method that:

type $g \in S(\square)$	$ \text{Fix } g $
(a)	3^6
(b)	3^3
(c)	3^4
(d)	3^2
(e)	3^3

so that the number of orbits is:

$$\frac{1}{24}(3^6 + 6 \cdot 3^3 + 3 \cdot 3^4 + 8 \cdot 3^2 + 6 \cdot 3^3) = 57 \quad (20.4.19)$$

hence there are 57 different colored cubes. ◀

Sylow theorems

21

Rings

22

Polynomials

23

Modules

24

Part III

Representation Theory and Lie Algebra

Part IV

**Differential Equations (get the
PDE chapters working)**

Fundamentals

25.1 Definitions

Definition 23.1 (*n*th order ODE)

An *n*th order ordinary differential equation (ODE) in \mathbb{K}^N is an equation:

$$y^{(n)} = f(t, y, y' \dots y^{(n-1)}) \quad (25.1.1)$$

for $(t, y, y' \dots y^{(n-1)}) \in \Omega, t \in \mathbb{R}, y \in \mathbb{K}^n$ where $\Omega \subseteq \mathbb{R} \times (\mathbb{K}^N)^n$ and $f : \Omega \rightarrow \mathbb{K}^n$, with $n, N \in \mathbb{N}^*$.

We highlight the fundamental case where $n = 1$, in which case:

$$y' = f(t, y), \quad (t, y) \in \Omega, \quad t \in \mathbb{R}, \quad y \in \mathbb{K}^N \quad (25.1.2)$$

Definition 23.2 (*Solution*)

A solution to an *n*th order ODE consists of a function $y : I \rightarrow \mathbb{K}^N$ *n* times differentiable such that:

- (i) $\forall t \in I, (t, y, y' \dots y^{(n-1)}) \in \Omega$
- (ii) $\forall t \in I, y^{(n)} = f(t, y, y' \dots y^{(n-1)})$

Definition 23.3 (*Linear and Homogeneous*)

An ODE is said to be **linear** if f is a polynomial function, that is:

$$f(t, y, y' \dots y^{(n-1)}) = \sum_{i=0}^n A_i(t) y^{(i)} \quad (25.1.3)$$

where $A_i(t) \in \text{Mat}_N(\mathbb{K})$ for $i = 1, 2..n$ and $A_0(t) \in \mathbb{K}^N$. A linear ODE is further said to be **homogeneous** if $A_0(t) = 0$. Homogeneous solutions are invariant under scaling $(t, y, y' \dots y^{(n-1)}) \rightarrow (\lambda t, \lambda y, \lambda y' \dots \lambda y^{(n-1)})$ for $\lambda \in \mathbb{K}$.

Proposition 23.4 (Smoothness of Solutions)

If $f : \Omega \rightarrow \mathbb{K}^N$ is of class C^k then all solutions y of (23.0.2) are of class C^{k+1} .

Proof. We provide a proof by induction. For $k = 0$, then f is continuous everywhere, and let y be a solution. Then, we have that $y' = f(t, y)$ and so y is continuous and differentiable everywhere, hence of class C^1 .

Let us now suppose that proposition 23.4 is true for some $k \in \mathbb{N}$, let f be of class C^{k+1} and let y be a solution. Then, since f is also of class C^k , then by hypothesis y must be of class C^{k+1} . However, $y' = f(t, y)$ is of class C^{k+1} and y is therefore of class C^{k+2} . ■

25.2 Integral formulation

There is a remarkable relationship between differential equations and integral equations. In this course we will consider four main types of integral equations. Suppose that $f : [a, b] \rightarrow \mathbb{R}$ and $K : [a, b]^2 \rightarrow K$ are continuous, with $t \in [a, b]$ then:

$$\text{Volterra non-homogeneous : } y(t) = f(t) + \int_a^t K(t, s)y(s)ds \quad (25.2.1)$$

$$\text{Fredholm non-homogeneous : } y(t) = f(t) + \lambda \int_a^b K(t, s)y(s)ds \quad (25.2.2)$$

and the two corresponding homogeneous equations. We call $K(t, s)$ the **kernel** of the integral equation.

Note that for the Fredholm homogeneous equation:

$$y(t) = \lambda \int_a^b K(t, s)y(s)ds \quad (25.2.3)$$

we may consider this as an eigenfunction equation, with λ as an eigenvalue and y as an eigenfunction.

Lemma 1 Suppose that $f : [a, b] \rightarrow \mathbb{R}$ is continuous. Then:

$$\int_a^x \int_a^{x'} f(t)dt dx' = \int_a^x (x-t)f(t)dt \quad (25.2.4)$$

Proof. Define the integral transform $G : [a, b] \rightarrow \mathbb{R}$ by:

$$F(x) = \int_a^x (x-t)f(t)dt \quad (25.2.5)$$

then because $f(t)$ and $x-t$ are continuous we may use the Leibniz integral rule to find:

$$F'(x) = \underbrace{[(x-t)f(t)]}_{t=x} \xrightarrow{\frac{\partial}{dx}} (x) + \int_a^x \frac{\partial}{\partial x} [(x-t)f(t)]dt = \int_a^x f(t)dt \quad (25.2.6)$$

We then deduce from the fundamental theorem of Calculus that:

$$\int_a^{x'} F'(x)dx = F(x') - F(a) \stackrel{0}{=} \int_a^{x'} \int_a^x f(t)dtdx \quad (25.2.7)$$

Substituting $x \rightarrow x'$ then:

$$F(x) = \int_a^x \int_a^{x'} f(t)dtdx' \quad (25.2.8)$$

as we wished to show. ■

Consider the differential equation $y'' + \lambda y = g(t)$ with $t \in [0, L]$. The reader will probably be familiar already with the solution, but here we wish to find the equivalent integral equation.

The first step in doing so is integrating from 0 to x to find (using the fact that y, y'' must both be continuous):

$$y'(t) - y'(0) + \lambda \int_0^t y(s)ds = \int_0^t g(s)ds \quad (25.2.9)$$

Further integration gives:

$$y(t) - y(0) - ty'(0) + \lambda \int_0^t (t-s)f(s)ds = \int_0^t (t-s)g(s)ds \quad (25.2.10)$$

where we used Lemma 1 to simplify the two double integrals.

We must now set some conditions to solve the problem explicitly.

Definition 23.5 (Initial and Boundary conditions)

An *initial condition* is a specification of $(t, y, y', \dots, y^{(n)})$ for n initial values $t = t_i$.

A *boundary condition* is a specification of y at the end-points of an interval.

1. **Initial condition:** suppose $y(0) = 0$ and $y'(0) = A$. Then:

$$y(t) = At + \int_0^t (t-s)g(s)ds - \lambda \int_0^t (t-s)y(s)ds \quad (25.2.11)$$

which is a Volterra non-homogeneous integral equation with $K(t, s) = \lambda(t-s)$ and $f(t) = At + \int_0^t (t-s)g(s)ds$.

2. **Boundary condition:** suppose $y(0) = 0$ and $y(L) = B$. Then we find upon inserting $t = L$ that:

$$y'(0) = \frac{1}{L} \left(\lambda \int_0^L (L-s)y(s)ds - \int_0^L (L-s)g(s)ds + B \right) \quad (25.2.12)$$

and substituting back into the original integral equation we find:

$$y = \frac{Bt}{L} - \int_0^L \frac{t}{L}(L-s)g(s)ds + \int_0^t (t-s)g(s)ds + \lambda \left(\int_0^L \frac{t}{L}(L-s)g(s)ds - \int_0^t (t-s)g(s)ds \right) \quad (25.2.13)$$

If we now define a function $K(s, t)$ such that:

$$\int_0^L K(s, t)f(s)ds = \int_0^L \frac{t}{L}(L-s)f(s)ds - \int_0^t (t-s)f(s)ds \quad (25.2.14)$$

then we may write:

$$y = \underbrace{\frac{Bt}{L} - \int_0^L K(s, t)g(s)ds}_{f(x)} + \lambda \left(\int_0^L K(s, t)y(s)ds \right) \quad (25.2.15)$$

It turns out that the kernel is¹ :

$$K(s, t) = \begin{cases} \frac{s}{L}(L-t) & \text{when } 0 \leq s \leq t \leq L \\ \frac{t}{L}(L-s) & \text{when } 0 \leq t \leq s \leq L \end{cases} \quad (25.2.20)$$

We therefore have a non-homogeneous Fredholm equation.

It is clear that the set of conditions we impose also affects the form of the equivalent integral equation. An ODE by itself without initial/boundary conditions is not enough.

This is because an ODE by itself can't be solved exactly, that is, we cannot find a particular solution, just a general solution. An integral equation however has an exact solution, with no free parameters, and can't therefore be associated to an ODE alone.

25.3 Picard iteration

Consider the Volterra integral equation:

$$y(t) = f(t) + \int_a^t K(s, t)y(s)ds \quad (25.3.1)$$

¹Indeed:

$$\int_0^L K(s, t)f(s)ds = \int_0^t K(s, t)f(s)ds + \int_t^L K(s, t)f(s)ds \quad (25.2.16)$$

$$= \int_0^t \frac{s}{L}(L-t)f(s)ds + \int_t^L \frac{t}{L}(L-s)f(s)ds \quad (25.2.17)$$

$$= \int_0^t \frac{s}{L}(L-t)f(s)ds + \int_0^L \frac{t}{L}(L-s)f(s)ds - \int_0^t \frac{t}{L}(L-s)f(s)ds \quad (25.2.18)$$

$$= \int_0^t (s-t)f(s)ds + \int_0^L \frac{t}{L}(L-s)f(s)ds \quad (25.2.19)$$

as required.

with f continuous on $[a, b]$ and $K, \partial_x K$ continuous on $[a, b]^2$. Our goal will be to define an iterative sequence (y_n) which improves as n increases. We should therefore make an initial guess, and insert that into the equation to find a better solution. We will define this sequence, known as a **Picard iteration** as follows:

$$\begin{cases} y_0 = f(t) \\ y_k(t) = f(t) + \int_a^t K(s, t)y_{k-1}(s)ds \end{cases} \quad (25.3.2)$$

Because $f(t)$ is continuous by hypothesis, so are all y_i for $i = 0, 1, 2, \dots$. We now conjecture that:

Proposition 23.6 (Picard iteration convergence)

We have that:

$$|u_n(t)| = |y_n(t) - y_{n-1}(t)| \leq M_n \quad (25.3.3)$$

with $\sum_{n=1}^{\infty} M_n$ converging.

Proof. Since K, f are continuous over $[a, b]$, they must be bounded:

$$|K(s, t)| \leq L, |f(t)| \leq M \quad \forall s, t \in [a, b] \quad (25.3.4)$$

We can therefore write:

$$|y_1(t) - y_0(t)| = \left| \int_a^t K(s, t)y_0(s)ds \right| = \int_a^x |K(s, t)||f(s)|ds \leq LM(x - a) \quad (25.3.5)$$

Let us now suppose that for some $n \geq 2$:

$$|y_{n-1}(t) - y_{n-2}(t)| \leq L^{n-1}M \frac{(t-a)^{n-1}}{(n-1)!} \quad (25.3.6)$$

We then find:

$$|y_n(s) - y_{n-1}(s)| = \left| \int_a^t K(s, t)(y_{n-1}(s) - y_{n-2}(s))ds \right| \quad (25.3.7)$$

$$\leq \int_a^t |K(s, t)||y_{n-1}(s) - y_{n-2}(s)|ds \quad (25.3.8)$$

$$\leq \int_a^t L^n M \frac{(s-a)^{n-1}}{(n-1)!} ds \quad (25.3.9)$$

$$= \leq L^n M \frac{(t-a)^n}{n!} \quad (25.3.10)$$

as required.

We can therefore define M_n as:

$$|y_n(t) - y_{n-1}(t)| \leq L^n M \frac{(t-a)^n}{n!} \leq L^n M \frac{(b-a)^n}{n!} \equiv M_n \quad (25.3.11)$$

Consequently:

$$\sum_{n=1}^{\infty} M_n = M(e^{L(b-a)} - 1) \quad (25.3.12)$$

and we then find that $\sum_{n=1}^{\infty} (y_n - y_{n-1})$ converges uniformly to u on $[a, b]$ using the Weierstrass test. \blacksquare

Notice however that this is a telescopic sum equal to $y - y_0 = u$ which implies that $y = u + y_0$.

We can then assert that $\forall \epsilon > 0, \exists N$ with:

$$|y(x) - y_n(x)| < \epsilon \quad \forall n \geq N \quad (25.3.13)$$

This implies that:

$$|K(s, t)y(s) - K(s, t)y_n(s)| < L\epsilon \quad \forall n \geq N \quad (25.3.14)$$

This is equivalent to saying that our iteration converges to the non-homogeneous Fredholm equation:

$$\int_a^t K(s, t)y_n(s)ds \longrightarrow \int_a^t K(s, t)y(s)ds, \quad \text{as } n \rightarrow \infty \quad (25.3.15)$$

We have therefore shown the existence of a continuous solution, but what about its uniqueness?

Suppose there is another solution Y so that:

$$|y(t) - Y(t)| \leq P \quad (25.3.16)$$

Let us suppose inductively that:

$$|y(t) - Y(t)| \leq L^{n-1} \frac{(t-a)^{n-1}}{(n-1)!} \quad (25.3.17)$$

Then:

$$|y(t) - Y(t)| = \left| \int_a^t K(s, t)(y(s) - Y(s))ds \right| \quad (25.3.18)$$

$$\leq L^n P \frac{(t-a)^n}{n!} \quad (25.3.19)$$

$$\leq L^n P \frac{(b-a)^n}{n!} \quad (25.3.20)$$

As $n \rightarrow \infty$ the RHS tends to zero, and we therefore find that $y = Y$ thus proving the uniqueness.

One can use a very similar process to the Fredholm equation using the iteration:

$$\begin{cases} y_0 = f(t) \\ y_k(t) = f(t) + \lambda \int_a^t K(s, t) y_{k-1}(s) ds \end{cases} \quad (25.3.21)$$

In this case we find that:

$$|y_n(t) - y_{n-1}(t)| \leq |\lambda|^n L^n M(b-a)^n \quad (25.3.22)$$

is uniformly convergent only if $|\lambda| \leq \frac{1}{L(b-a)}$. This is the sufficient condition that must be met for a solution to exist.

25.4 Existence and uniqueness

Consider the Cauchy problem:

$$y' = f(x, y), \quad y(a) = c \quad (25.4.1)$$

where f satisfies the following two conditions:

- (i) f is continuous in a region U containing $R = \{(x, y) : |x - a| \leq h, |y - c| \leq k\} \subseteq U$.
- (ii) f satisfies the Lipschitz condition:

$$|f(x, y_1) - f(x, y_2)| \leq A|y_1 - y_2|, \quad \forall (x, y_1), (x, y_2) \in U \quad (25.4.2)$$

- (iii) Defining:

$$M = \sup\{|f(x, y)| : (x, y) \in R\} \quad (25.4.3)$$

then we require:

$$Mh \leq k \quad (25.4.4)$$

If these three conditions are satisfied, a very important result, known as the Cauchy-Picard existence and uniqueness theorem, is established.

Theorem (Cauchy-Picard Existence and Uniqueness theorem)

If (i), (ii), (iii) are all satisfied then there exists for $|x - a| \leq h$ a solution to the Cauchy problem:

$$y' = f(x, y), \quad y(a) = c \quad (25.4.5)$$

and this solution is unique in U .

We will present the proof of a more general result later.

First Order ODEs

Theorem 1 (Existence and Uniqueness)

Let $f(t, y)$ and $\frac{\partial f}{\partial y}$ exist and be continuous on some domain $\mathcal{D} \subset \mathbb{R}^2$. Then: $\forall (t_0, y_0) \in \mathcal{D}, \exists P t$ such that the Cauchy problem:

$$\begin{cases} \dot{y} = f(t, y) \\ y(t_0) = y_0 \end{cases}$$

has a unique solution in the interval $I = [t_0 - Pt, t_0 + Pt]$. If $y_1(t)$ and $y_2(t)$ are both solutions on I_1 and I_2 respectively, then $y_1(t) = y_2(t)$, that is, a the solution is unique.

26.1 Exact Differential Equations

Consider an ODE in full differentials, with solutions $\Phi(x, y)$ such that:

$$\begin{cases} \forall (x, y) \in \mathcal{D}, \Phi_x = P(x, y), \Phi_y = Q(x, y) \\ d\Phi = 0 \end{cases}$$

We can then rewrite, using the chain rule, the equation as:

$$P(x, y)dx + Q(x, y)dy = 0 \quad (26.1.1)$$

Theorem. If $\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}$ through a simply connected domain \mathcal{D} , then $Pdx + Qdy = 0$ is an exact differential equation.

Proof. Consider the solution $\Phi(x, y) = C$, for some constant C. It follows from the chain rule that:

$$\frac{\partial \Phi}{\partial x} = P, \frac{\partial \Phi}{\partial y} = Q \implies \frac{\partial^2 \Phi}{\partial x \partial y} = \frac{\partial^2 \Phi}{\partial y \partial x} \implies \frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x} \blacksquare$$

Strategy 24.1 () To solve:

1. Set the equation: $\Phi_x = P(x, y)$, and integrate directly with respect to x:

$$\Phi = \int P(x, y)dx + \phi(y)$$

2. Substitute this into $\Phi_y = Q(x, y)$:

$$\frac{d}{dy} \left(\int P(x, y)dx + \phi(y) \right) = Q(x, y)$$

and rearrange to find:

$$\phi(y) = \int \left(Q(x, y) - \frac{d}{dy} \int P(x, y)dx \right) dy$$

3. Set:

$$\int P(x, y)dx + \int \left(Q(x, y) - \frac{d}{dy} \int P(x, y)dx \right) dy = C$$

for some constant C .

26.2 Separable Differential Equations

In the case where $P(x, y)$ is independent of y , and $Q(x, y)$ is independent of x in (2.1), then we have the separable differential equation:

$$P(x)dx + Q(y)dy = 0 \quad (26.2.1)$$

To solve, integrate directly:

$$\int P(x)dx + \int Q(y)dy = 0$$

and simplify.

26.3 Inexact Differential Equations

Inexact differential equations are of the form:

$$\begin{cases} P(x, y)dx + Q(x, y)dy = 0 \\ \frac{\partial P}{\partial y} \neq \frac{\partial Q}{\partial x} \end{cases} \quad (26.3.1)$$

unlike exact differential equations. Firstly, consider an ODE of the following form:

$$\frac{dy}{dx} = f(ax + by). \quad (26.3.2)$$

To solve:

1. Apply the change of variables $z(x) = ax + by(x)$ to find:

$$\frac{dz}{dx} = a + b\frac{dy}{dx}$$

2. Substitute the expression for y' :

$$\frac{dz}{dx} = a + bf(z)$$

3. Solve as a separable differential equation:

$$\int \frac{dz}{a + bf(z)} = \int dx$$

Next, consider the case where (2.3) is homogeneous, that is:

$$\frac{P(\lambda x, \lambda y)}{P(x, y)} = \frac{Q(\lambda x, \lambda y)}{Q(x, y)}, \quad \forall \lambda \neq 0$$

Then:

$$\frac{dy}{dx} = -\frac{P(x, y)}{Q(x, y)} = -\frac{P(\lambda x, \lambda y)}{Q(\lambda x, \lambda y)}|_{\lambda=\frac{1}{x}} = -\frac{P(1, \frac{y}{x})}{Q(1, \frac{y}{x})} = f\left(\frac{y}{x}\right)$$

To solve:

1. Apply the change of variables $u(x)x = y(x)$:

$$x\frac{du}{dx} + u = \frac{dy}{dx}$$

2. Substitute into original ODE:

$$f(u) = u + x\frac{du}{dx}$$

and separate variables:

$$\int \frac{du}{f(u) - u} = \int \frac{dx}{x}.$$

where $f(u) \neq u$.

26.4 Integrating Factor Method

Consider the non homogeneous ODE of the form:

$$\dot{y}(t) = a(t)y(t) + f(t) \quad (26.4.1)$$

To solve, we wish to multiply the whole equation by a so called *integrating factor* Λ such that $\dot{\Lambda} = \Lambda a(t)$. Then:

$$\begin{aligned} \Lambda \dot{y}(t) - \overbrace{\Lambda a(t)}^{\dot{\Lambda}} y(t) &= \Lambda f(t) \\ \frac{d\Lambda y(t)}{dt} &= \Lambda f(t) \\ \implies y(t) &= \frac{1}{\Lambda(t)} \left[C + \int_0^t \Lambda(t') f(t') dt' \right] \end{aligned}$$

To find the integrating factor:

$$\begin{aligned} \frac{d\Lambda}{dt} &= \Lambda(t)a(t) \\ \therefore \int \frac{d\Lambda}{\Lambda} &= \int a(t) dt \\ \therefore \Lambda(t) &= \exp \left(\int_0^t a(t') dt' \right) \end{aligned}$$

26.5 Bernoulli Equations

Finally, let us look at the Bernoulli Equations:

$$y' + a(x)y = b(x)y^n$$

which due to the y^n term, is non linear.

To solve:

1. Divide through by y^n to find:

$$\frac{dy}{dx} y^{-n} + a(x) y^{1-n} = b(x)$$

2. Apply a change of variables $u = y^{1-n}$:

$$\frac{du}{dx} + (1-n)a(x)u = (1-n)b(x)$$

3. Use integrating factor method.

26.6 Stability and Equilibrium points

Definition 2.1. An *equilibrium point* of a differential equation is a constant solution $y' = 0, \forall t \in \mathcal{D}$. It is:

Stable: if $y \rightarrow c$ as $t \rightarrow \infty$, in other words the deviation decays.

Unstable: if $y \rightarrow \infty$ as $t \rightarrow \infty$, in other words the deviation grows.

We can linearize differential equations by doing perturbation analysis. Suppose $y = a$ is an equilibrium point of $y' = f(x, y)$. We induce an arbitrarily small perturbation $y = a + \epsilon(t)$, so that:

$$\begin{aligned}\frac{d\epsilon}{dt} &= \frac{dy}{dt} = f(a, t) + \epsilon \frac{\partial f}{\partial y}(a, t) + O(\epsilon^2) \\ &\approx \epsilon \frac{\partial f}{\partial t}(a, t)\end{aligned}$$

If $\dot{\epsilon} > 0$, we have unstable equilibrium, if $\dot{\epsilon}(t) < 0$, we have stable equilibrium.

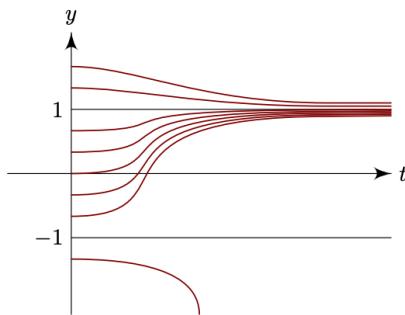


Figure 26.1. Integral curves for $\dot{y} = t(1 - y^2)$ [?]

Second Order ODEs

27.1 Homogeneous equation

We consider the second order homogeneous differential equation with initial conditions:

$$\begin{cases} a(x)y'' + b(x)y' + c(x)y = 0, \quad (x \in [a, b]) \\ y(x_0) = y_0 \\ y'(x_0) = y'_0 \end{cases} \quad (27.1.1)$$

where $a(x) > 0, \forall x \in [a, b]$.

Definition (Fundamental matrix and Wronskian) Suppose $y_1(x)$ and $y_2(x)$ are differentiable functions on $[a, b]$. We then define their **Fundamental matrix** to be:

$$\mathbf{Y}(y_1, y_2) = \begin{pmatrix} y_1(x) & y_2(x) \\ y'_1(x) & y'_2(x) \end{pmatrix} \quad (27.1.2)$$

and define their *Wronskian* to be the determinant of the fundamental matrix:

$$W(y_1, y_2)(x) = \begin{vmatrix} y_1(x) & y_2(x) \\ y'_1(x) & y'_2(x) \end{vmatrix} = y_1(x)y'_2(x) - y'_1(x)y_2(x) \quad (27.1.3)$$

Proposition (Linear (in)dependence)

For two non-constant functions $y_1(x), y_2(x)$ differentiable on $[a, b]$:

- a) If the Wronskian $W(y_1, y_2)(x_0) \neq 0$ for some $x_0 \in [a, b]$, the two functions $y_1(x)$ and $y_2(x)$ are linearly independent on $[a, b]$.
- b) If they are linearly dependent then $W(y_1, y_2)(x) = 0, \forall x \in [a, b]$.

Proof.

- a. Assume that the Wronskian is non-zero for some $x_0 \in [a, b]$, then:

$$y_1(x_0)y'_2(x_0) - y_2(x_0)y'_1(x_0) \neq 0 \implies y_1(x_0) \neq \frac{y'_1(x_0)}{y'_2(x_0)}y_2(x_0) = cy_1(x_0) \quad (27.1.4)$$

where we set $c = \frac{y'_1(x_0)}{y'_2(x_0)}$. Therefore $y_1(x_0)$ and $y_2(x_0)$ are linearly independent. Using the result b) we see that $y_1(x)$ and $y_2(x)$ must be linearly independent for all $x \in [a, b]$, since otherwise the Wronskian would vanish identically.

- b. Assume that the solutions are linearly dependent, so that: $c_1 y_1(x) + c_2 y_2(x) = 0$ for some $c_1, c_2 \in \mathbb{R}$ not both zero. Then differentiating with respect to x one finds that:

$$\begin{cases} c_1 y_1(x) + c_2 y_2(x) = 0 \\ c_1 y'_1(x) + c_2 y'_2(x) = 0 \end{cases} \quad (27.1.5)$$

which we may consider as a system of equations in c_1, c_2 . For a non-zero solution (c_1, c_2) to exist the following determinant must vanish:

$$\begin{vmatrix} y_1(x) & y_2(x) \\ y'_1(x) & y'_2(x) \end{vmatrix} = 0 \implies W(y_1, y_2)(x) = 0, \forall x \in [a, b] \quad (27.1.6)$$

thus proving the desired result. as required. ■

Suppose that by some stroke of luck we have already found two linearly independent solutions, $y_1(x)$ and $y_2(x)$, of (27.1.1). Due to the linearity of (27.1.1), we may use the principle of superposition and state that:

$$y(x) = c_1 y_1(x) + c_2 y_2(x), \forall c_1, c_2 \in \mathbb{R}$$

will also be a solution. We now determine the constants:

$$\begin{pmatrix} y(x_0) \\ y'(x_0) \end{pmatrix} = \begin{pmatrix} y_1(x_0) & y_2(x_0) \\ y'_1(x_0) & y'_2(x_0) \end{pmatrix} \cdot \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \mathbf{Y} \cdot \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

Then, the coefficients are determined as:

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} y_0 \\ y'_0 \end{pmatrix} \cdot \mathbf{Y}^{-1} \quad (27.1.7)$$

$$= \frac{1}{W(y_1, y_2)(x_0)} \begin{pmatrix} y'_2(x_0) & -y_2(x_0) \\ -y'_1(x_0) & y_1(x_0) \end{pmatrix} \cdot \begin{pmatrix} y(x_0) \\ y'(x_0) \end{pmatrix} \quad (27.1.8)$$

Note that had we chosen linearly dependent solutions then \mathbf{Y} would not have been invertible, and thus we would not be able to find c_1, c_2 directly through this method.

Theorem (Abel's Identity for second order ODE)

If $y_1(x)$ and $y_2(x)$ are solutions to the homogeneous ODE:

$$y'' + p(x)y' + q(x)y = 0 \quad (27.1.9)$$

then:

$$W(y_1, y_2)(x) = W(y_1, y_2)(x_0) \exp \left[- \int_{t_0}^x p(s) ds \right] \quad (27.1.10)$$

Proof. The derivative of a determinant is given by Jacobi's formula:

$$\frac{d}{dx} \det\{\mathbf{A}(x)\} = \text{tr} \left(\text{adj } \mathbf{A}(x) \frac{d\mathbf{A}(x)}{dx} \right) \quad (27.1.11)$$

For a 2×2 matrix \mathbf{A} :

$$\mathbf{A} = \begin{pmatrix} a(x) & b(x) \\ c(x) & d(x) \end{pmatrix} \implies \text{adj } \mathbf{A} = \begin{pmatrix} d(x) & -b(x) \\ -c(x) & a(x) \end{pmatrix} \quad (27.1.12)$$

so that:

$$\text{adj } \mathbf{A}(x) \frac{d\mathbf{A}(x)}{dx} = \begin{pmatrix} d(x) & -b(x) \\ -c(x) & a(x) \end{pmatrix} \begin{pmatrix} a'(x) & b'(x) \\ c'(x) & d'(x) \end{pmatrix} \quad (27.1.13)$$

$$= \begin{pmatrix} a'(x)d(x) - b(x)c'(x) & b'(x)d(x) - b(x)d'(x) \\ a(x)c'(x) - a'(x)c(x) & a(x)d'(x) - c(x)b'(x) \end{pmatrix} \quad (27.1.14)$$

and hence:

$$\frac{d}{dx} \det\{\mathbf{A}\} = a'(x)d(x) - b(x)c'(x) + a(x)d'(x) - c(x)b'(x) \quad (27.1.15)$$

Therefore, substituting $a(x) = y_1(x)$, $c(x) = y'_1(x)$, $b(x) = y_2(x)$, $d(x) = y'_2(x)$ we find that:

$$W'(y_1, y_2)(x) = y'_1(x)y'_2(x) - y_2(x)y''_1(x) + y_1(x)y''_2(x) - y'_1(x)y'_2(x) \quad (27.1.16)$$

$$= y_1(x)y''_2(x) - y_2(x)y''_1(x) \quad (27.1.17)$$

We may now substitute $y''_i(x) + p_i(x)y'_i(x) + q_i(x) = 0$ for $i = 1, 2$ to find that

$$\begin{aligned} W'(y_1, y_2)(x) &= y_1(x)(-p(x)y'_2(t) - q(x)y_2(x)) - y_2(x)(-p(x)y'_1(x) - q(x)y_1(x)) \\ &= -(y_1(x)y'_2(x) - y_2(x)y'_1(x))p(x) \\ &= -W(y_1, y_2)(x)p(x) \end{aligned}$$

Separating variables, and integrating from x_0 to x , one finally finds:

$$W(y_1, y_2)(x) = W(y_1, y_2)(x_0) \exp \left[- \int_{x_0}^x p(s) ds \right]$$

as required. ■

27.2 Non-homogeneous

Definition 3.4. The second order non-homogeneous differential equation is:

$$\begin{cases} a(t)y'' + b(t)y' + c(t)y = f(t) \\ y(t_0) = y_0 \\ y'(t_0) = y'_0 \end{cases} \quad (27.2.1)$$

and its *associated homogeneous equation* is (3.1).

Theorem 3 (Complementary Function and Particular Integral) The general solution to the non-homogeneous differential equation may be written as the sum:

$$y(t) = y_{CF}(t) + y_{PI}(t)$$

where y_{CF} is the complementary function, that is, the solution to the associated homogeneous equation, and y_{PI} is a particular solution.

Proof.

Consider the difference between the general solution and the particular integral: $y(t) - y_{PI}(t)$. This must be a solution to the associated homogeneous equation by the superposition principle. Indeed:

$$\begin{aligned} & a(t)(y''(t) - y''_{PI}(t)) + b(t)(y'(t) - y'_{PI}(t)) + c(t)(y(t) - y_{PI}(t)) \\ &= a(t)y''(t) + b(t)y'(t) + c(t)y(t) - (a(t)y''_{PI}(t) + b(t)y'_{PI}(t) + c(t)y_{PI}(t)) \\ &= f(t) - f(t) = 0 \end{aligned}$$

as required. Hence, since $y_1(t)$ and $y_2(t)$ form a fundamental set of solutions, we may write any solution of the associated homogeneous equation, including $y(t) - y_{PI}(t)$, as a linear combination:

$$y(t) - y_{PI}(t) = c_1 y_1(t) + c_2 y_2(t) \implies y(t) = y_{PI}(t) + c_1 y_1(t) + c_2 y_2(t)$$

as required. ■

It may seem like this theorem is of very little use, since a particular solution is *ipso facto* given by the general solution. However, the following two methods may be used to determine the particular solution:

1. Undetermined Coefficients: guessing and checking, quick but works only in special cases.
2. Variation of parameters: more general, almost always works.

27.3 Undetermined Coefficients

This method consists in guessing the form of the particular integral leaving the coefficients indeterminate, and then plug them into the differential equation.

The following table summarizes possible particular integrals for different functions $f(t)$:

$g(t)$	$y_{PI}(t)$
$\alpha e^{\beta t}$	$Ae^{\beta t}$
$a \cos(\beta t) + b \sin(\beta t)$	$A \cos(\beta t) + B \sin(\beta t)$
$\sum_{i=0}^n a_i x^i$	$\sum_{i=0}^n A_i x^i$

27.4 Variation of Constants

Assume that we have found the complementary solution to the non-homogeneous equation:

$$y_{CF}(t) = c_1 y_1(t) + c_2 y_2(t)$$

and look for the particular integral of the form:

$$y_{PI}(t) = \psi_1 y_1 + \psi_2 y_2 \quad (27.4.1)$$

such that:

$$\psi'_1 y_1 + \psi'_2 y_2 = 0$$

Differentiating:

$$\begin{aligned} y'_{PI} &= \psi_1 y'_1 + \psi_2 y'_2 \\ y''_{PI}(t) &= \psi'_1 y'_1 + \psi'_2 y'_2 + \psi_1 y''_1 + \psi_2 y''_2 \end{aligned}$$

Inserting these into the differential equation and simplifying:

$$\psi'_1 y'_1 + \psi'_2 y'_2 = \frac{f(t)}{a(t)}$$

Let us finally assume that $a(t) = 1$ (which corresponds to rearranging the equation so that the coefficient of y'' is 1), then:

$$\begin{cases} \psi'_1 y'_1 + \psi'_2 y'_2 = f(t) \\ \psi'_1 y_1 + \psi'_2 y_2 = 0 \end{cases}$$

$$\therefore \psi'_1 = -\frac{\psi'_2 y_2}{y_1} \implies -\frac{\psi'_2 y_2}{y_1} y'_1 + \psi'_2 y'_2 = f(t)$$

$$\psi'_2 = \frac{y_1 f(t)}{y_1 y'_2 - y_2 y'_1}$$

Since y_1 and y_2 are linearly independent, their Wronskian is non-zero, and thus:

$$\psi'_1 = -\frac{y_1 f(t)}{y_2 y'_2 - y_2 y'_1}, \quad \psi'_2 = \frac{y_1 f(t)}{W(y_1, y_2)},$$

which can be integrated directly in (3.4) to get:

$$y_{PI}(t) = -y_1 \int \frac{y_1 f(t)}{y_2 y'_2 - y_2 y'_1} dt + y_2 \int \frac{y_1 f(t)}{W(y_1, y_2)} dt \quad (27.4.2)$$

27.5 Reduction of Order

Finally, let us look at how we may simplify differential equations when one solution to the associated homogeneous equation, $y_1(t)$ is known.

Then, to solve:

1. We search for solutions of the form:

$$y(t) = \phi(t)y_1(t)$$

and evaluate its first and second derivatives.

2. We substitute into the original differential equation.
3. Solve the resulting differential equation by substituting $\psi = \phi'$.

Mechanical Vibrations and Resonance Phenomena

Constant coefficient second-order linear ODEs are of particular interest in the area of mechanical vibrations.

28.1 Homogeneous Equation

Consider the homogeneous equation:

$$\ddot{y} + a\dot{y} + by = 0 \quad (28.1.1)$$

and we guess a solution in exponential form:

$$y(t) = Ce^{\lambda t}$$

Plugging into (4.1) one finds:

$$\lambda^2 + a\lambda + b = 0 \quad (28.1.2)$$

which is called the *auxiliary equation*. The quadratic formula then yields:

$$\lambda_{1,2} = \frac{-a \pm \sqrt{a^2 - 4b}}{2}$$

which brings us to the following result.

Proposition 1

The solutions to the second order homogeneous ODE with constant coefficients is:

$$y(t) = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t}$$

provided that the solutions λ_1, λ_2 are non-degenerate. In the case where:

$a^2 > 4b$: we have two real, distinct solutions, and by the superposition principle:

$$y(t) = e^{-at/2} (C_1 e^{t\sqrt{a^2 - 4b}/2} + C_2 e^{-t\sqrt{a^2 - 4b}/2}) \quad (28.1.3)$$

$a^2 < 4b$: we have two complex, distinct solutions, and by the superposition principle:

$$y(t) = e^{-at/2} (C_1 e^{it\sqrt{4b-a^2}/2} + C_2 e^{-it\sqrt{4b-a^2}/2})$$

Euler's identity then gives:

$$y(t) = e^{-at/2} (A \cos \Omega t + B \sin \Omega t) = \alpha e^{-at/2} \cos(\Omega t + \phi) \quad (28.1.4)$$

$$\text{where } \Omega = \frac{\sqrt{4b-a^2}}{2}.$$

If instead we have two degenerate solutions, so that $a^2 = 4b$, then we have only found one solution:

$$y_1(t) = Ce^{-at/2}.$$

We use the method of reduction of variables to find the general solution:

$$\begin{aligned} y(t) &= \psi(t)e^{-at/2} \implies \psi'' + \left(b - \frac{a^2}{4}\right)\psi = 0 \\ \therefore \psi'' &= 0 \implies \psi = C_1 t + C_2 \end{aligned}$$

since $b = \frac{a^2}{4}$. Finally:

$$y(t) = (C_1 t + C_2)e^{-at/2} \quad (28.1.5)$$

28.2 Damped Harmonic Motion

For a damped harmonic system, there will be two forces in action, a restoring force $F_{res} = -m\omega_0^2 y$ and a damping force $F_{damp} = -m\gamma\dot{y}$. Newton's second law yields the second order ODE with constant coefficients:

$$\ddot{y} + \gamma\dot{y} + \omega_0^2 y = 0 \quad (28.2.1)$$

In this case, we may define:

$$\Omega \equiv \frac{\sqrt{\omega_0^2 - \frac{\gamma^2}{4}}}{2}$$

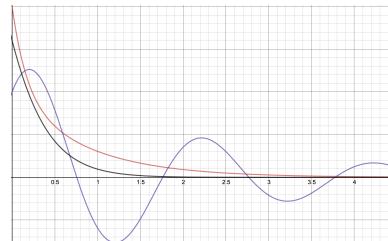


Figure 28.1. Plots of over damped (red), critically damped (black) and under damped (purple) solutions.

We may further impose the initial conditions $y(0) = y_0, \dot{y}(0) = 0$. As in the previous section, we consider three special cases:

$\boxed{\omega_0 > \lambda/2}$, the oscillator is *under damped*. The solution is:

$$y(t) = y_0 e^{-\gamma t/2} \left(\cos \Omega t + \frac{\gamma}{2\Omega} \sin \Omega t \right) \quad (28.2.2)$$

Note that this can be rewritten as:

$$y(t) = A e^{-\gamma t/2} \cos(\Omega t + \phi)$$

which implies that the period of oscillations is:

$$\tau = \frac{2\pi}{\Omega}$$

and over a cycle the amplitude is multiplied by:

$$e^{-\gamma\tau/2} = \exp \left(-\frac{2\pi\gamma}{\Omega} \right)$$

called the *amplitude decay factor*.

$\boxed{\omega_0 < \lambda/2}$, the oscillator is *over damped*. The solution is:

$$y(t) = \frac{y_0}{2\Omega} e^{-\gamma t/2} \left[\left(\Omega + \frac{\gamma}{2} \right) e^{\Omega t} + \left(\Omega - \frac{\gamma}{2} \right) e^{-\Omega t} \right] \quad (28.2.3)$$

$\boxed{\omega_0 = \lambda/2}$, the oscillator is *critically damped*, the solution is:

$$y(t) = y_0 e^{-\gamma t/2} \left(1 + \frac{\gamma}{2} t \right) \quad (28.2.4)$$

Definition 4.1 For a lightly damped oscillator in the regime $\gamma \ll \omega_0$, with initial stored energy E_0 and energy lost per period of oscillation PE_τ , the *quality factor* is defined as:

$$Q = \frac{2\pi E_0}{PE_\tau}$$

Proposition 2. *The quality factor for a damped oscillator is:*

$$Q = \frac{\omega_0}{\gamma}$$

Proof. The energy stored in the oscillator is:

$$E_0 = \frac{1}{2}ky_0^2 = \frac{1}{2}m\omega_0^2y_0^2$$

The under damped solution gives:

$$\begin{cases} y(t) = y_0 e^{-\gamma t/2} \cos(\omega_0 t - \phi) \\ \dot{y}(t) = -y_0 e^{-\gamma t/2} \left(\frac{\gamma}{2} \cos(\omega_0 t - \phi) + \omega_0 \sin(\omega_0 t - \phi) \right) \end{cases}$$

We then define the energy at t to be the sum of the kinetic and potential energy:

$$\begin{aligned} E(t) &= \frac{1}{2}m\dot{y}^2 + \frac{1}{2}m\omega_0 y^2 \\ \implies \frac{dE(t)}{dt} &= m\ddot{y}\dot{y} + m\omega_0 y\dot{y} = m\dot{y} \underbrace{(\ddot{y} + \omega_0 y)}_{-\omega_0^2 y} = -m\gamma\dot{y}^2 \\ \therefore PE_\tau &= \int_t^{t+Pt} -m\gamma\dot{y}^2 dt \\ &= \int_t^{t+Pt} m\gamma y_0^2 e^{-\gamma t} \left(\frac{\gamma}{2} \cos(\omega_0 t - \phi) + \omega_0 \sin(\omega_0 t - \phi) \right)^2 dt \end{aligned}$$

We can now apply the substitution $t' = \omega_0 t - \phi$, so that the limits of integration become 0 and 2π :

$$\begin{aligned} PE_\tau &= m\gamma y_0^2 \int_0^{2\pi} e^{-\gamma \frac{t'+\phi}{\omega_0}} \left(\frac{\gamma}{2} \cos t' + \omega_0 \sin t' \right)^2 \frac{1}{\omega_0} dt' \\ &= \frac{m\gamma y_0^2 \pi}{\omega_0} \frac{1}{4} (4\omega_0^2 + \gamma^2) \\ &= \pi m\gamma y_0^2 \omega_0 \left(1 + \frac{\gamma^2}{4\omega_0^2} \right) \\ &= \pi m\gamma y_0^2 \omega_0 \end{aligned}$$

hence:

$$Q = 2\pi \frac{\frac{1}{2}m\omega_0^2 y_0^2}{\pi m\gamma y_0^2 \omega_0} = \frac{\omega_0}{\gamma}$$

as required. ■

28.3 Forced Oscillations

Finally, let us consider the damped, forced oscillations equation:

$$\ddot{y} + \gamma\dot{y} + \omega_0^2 y = F \cos \omega t \quad (28.3.1)$$

We have already found the complementary function in the previous section, we must now find a particular solution. This can be done using the method of undetermined coefficients. To do so, we solve the complex version of (4.10):

$$\ddot{z} + \gamma\dot{z} + \omega_0^2 z = F e^{i\omega t}$$

and use as a trial solution $z = C e^{i\omega t}$. Then:

$$C = \frac{F}{\omega_0^2 - \omega^2 + i\omega\gamma} = \frac{F(\omega_0^2 - \omega^2 - i\omega\gamma)}{(\omega_0^2 - \omega^2)^2 + \omega^2\gamma^2}$$

Next, we define:

$$\cos \phi = \frac{\omega_0^2 - \omega^2}{\sqrt{(\omega_0^2 - \omega^2)^2 + \omega^2\gamma^2}}, \sin \phi = \frac{\omega\gamma}{\sqrt{(\omega_0^2 - \omega^2)^2 + \omega^2\gamma^2}}$$

so that:

$$C = \frac{F e^{-i\phi}}{\sqrt{(\omega_0^2 - \omega^2)^2 + \omega^2\gamma^2}}$$

This finally gives the solution:

$$z(t) = \frac{F e^{i(\omega_0 t - \phi)}}{\sqrt{(\omega_0^2 - \omega^2)^2 + \omega^2\gamma^2}}$$

Taking the real part yields:

$$y_{PI}(t) = A \cos(\omega t - \phi), \quad A = \frac{F}{\sqrt{(\omega_0^2 - \omega^2)^2 + \omega^2\gamma^2}} \quad (28.3.2)$$

Finally, the general solution is:

$$y(t) = \underbrace{A \cos(\omega t - \phi)}_{\text{steady state}} + \underbrace{e^{-\gamma t/2} [C_1 \cos \Omega t + C_2 \sin \Omega t]}_{\text{transient}}$$

where the first term is the steady state solution, and the second term is the transient solution, and quickly decays after $t \gg \gamma^{-1}$.

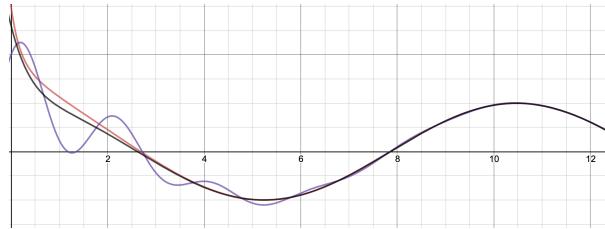


Figure 28.2. Forced solutions for over damped, under damped and critically damped oscillators.

28.4 Resonance

Let us now consider the scenario in which $\gamma \ll \omega_0$, then $A(\omega)$ has a peak near ω_0 , that is, when the frequency of the forced oscillations are close to the natural frequency of the oscillator. This phenomenon is known as *resonance*. To find the peak, we set $A'(\omega) = 0$:

$$4\omega_{res}(\omega_0 - \omega_{res}^2) + 2\omega_{res}\gamma^2 = 0 \implies \omega_{res} = \sqrt{\omega_0^2 - \frac{\gamma^2}{2}} \approx \omega_0$$

The peak amplitude is then:

$$A_{res} \approx A(\omega_0) = \frac{F}{\omega_0\gamma}$$

General Linear ODEs

29.1 Existence and Uniqueness

Theorem If all $A_{ij}(t)$ and $f_i(t)$ are continuous on $\mathcal{I} = (t_1, t_2)$, then $\forall t_0 \in \mathcal{I}, \forall \mathbf{y}_0 \in \mathbb{R}^n$, the Cauchy problem:

$$\begin{cases} \dot{\mathbf{y}} = A(t) \cdot \mathbf{y} + \mathbf{f}(t) \\ \mathbf{y}(t_0) = \mathbf{y}_0 \end{cases} \quad (29.1.1)$$

has a unique solution in \mathcal{I}

29.2 Fundamental set and Wronskians

Definition 5.1. The fundamental system of solutions of the homogeneous equation:

$$\dot{\mathbf{y}} = A(t) \cdot \mathbf{y}$$

is the set of linearly independent solutions to the associated homogeneous differential equation:

$$\{\mathbf{y}_1(t), \mathbf{y}_2(t), \dots, \mathbf{y}_n(t)\}$$

Lemma 2. The following are true for any Cauchy system of the form (3.1):

- a. If $\exists t_0 \in \mathcal{I}$ such that $\{\mathbf{y}_i(t_0)\}$ is linearly independent, then the fundamental set of solutions is linearly independent.
- b. If $\exists t_0 \in \mathcal{I}$ such that $\{\mathbf{y}_i(t_0)\}$ is linearly dependent, then the fundamental set of solutions is linearly dependent.

Proof.

- a. Suppose that $\{\mathbf{y}_i(t_0)\}$ is linearly dependent. Then, $\exists \{C_i\}$ not all equal to zero such that, $\forall t \in \mathcal{I}, C_i \mathbf{y}_i = \mathbf{0}$. However, this is not satisfied for $t = t_0$, thus we have a contradiction.
- b. Suppose that $\{\mathbf{y}_i(t_0)\}$ is linearly dependent. Then, $\forall t \in \mathcal{I}, \exists \{C_i\}$ such that: $\mathbf{y} = C_i \mathbf{y}_i = \mathbf{0}$. This implies that $\mathbf{y}(t_0) = \mathbf{0}$, and by the uniqueness theorem, $\mathbf{y}(t) = \mathbf{0}$ identically, and therefore $C_i \mathbf{y}_i(t) = (0), \forall t$ as required. ■

Definition 5.2. The fundamental set of solutions can be packed into a matrix, called the *fundamental matrix*:

$$\mathbf{Y}(t) = (\mathbf{y}_1(t) \ \mathbf{y}_2(t) \ \dots, \mathbf{y}_n(t)) \quad (29.2.1)$$

The *Wronskian* is then defined as the determinant of the fundamental matrix:

$$W(t) = \det \mathbf{Y}(t) \quad (29.2.2)$$

Theorem 3. (Liouville's formula) To compute the Wronskian, we can use Liouville's formula:

$$W(t) = W(t_0) \exp \left[\int_{t_0}^t \text{tr} A(t') dt' \right]$$

Proof. The derivative of the Wronskian is given by differentiating row by row and then summing:

$$\dot{W}(t) = \sum_{i=1}^n \det \{\mathbf{Y}_i\}^*(t)$$

where:

$$\mathbf{Y}_i^*(t) = \begin{pmatrix} y_{11} & y_{12} & \dots & y_{1n} \\ \vdots & \vdots & & \\ \dot{y}_{i1} & \dot{y}_{i2} & \dots & \dot{y}_{in} \\ \vdots & \vdots & & \\ y_{n1} & y_{n2} & \dots & y_{nn} \end{pmatrix} = \begin{pmatrix} y_{11} & y_{12} & \dots & y_{1n} \\ \vdots & \vdots & & \\ a_{ii}\dot{y}_{i1} & a_{ii}\dot{y}_{i2} & \dots & a_{ii}\dot{y}_{in} \\ \vdots & \vdots & & \\ y_{n1} & y_{n2} & \dots & y_{nn} \end{pmatrix}$$

This is because $\dot{Y}_{ik} = \sum_{j=1}^n A_{ij} Y_{jk}$, we can then subtract from the i th row the linear combination of all other rows:

$$\sum_{j \neq i}^n A_{ij} (\mathbf{Y}_{j1} \dots \mathbf{Y}_{jn})$$

and keep the determinant unchanged. Therefore:

$$\det \mathbf{Y}_i^*(t) = a_{ii} \det \mathbf{Y}(t) \implies \dot{W}(t) = \text{tr} A \cdot \mathbf{Y}(t)$$

which can be solved to yield the required solution. ■

29.3 Homogeneous ODE

Let us now look at how to solve the homogeneous ODE:

$$\begin{cases} \dot{\mathbf{y}}(t) = A(t) \cdot \mathbf{y} \\ \mathbf{y}(t_0) = \mathbf{y}_0 \end{cases}$$

If we know a fundamental system of solutions $\mathbf{y}_i(t)$, then by the principle of superposition, seeing as the elements of this set form a basis for all solutions, one finds that:

$$\mathbf{y}(t) = \mathbf{Y}(t) \cdot \mathbf{C}$$

We impose the condition $\mathbf{u}(t_0) = \mathbf{y}_0$ to get:

$$\mathbf{C} = \mathbf{Y}^{-1}(t_0) \cdot \mathbf{y}_0$$

where $\exists \mathbf{Y}^{-1}(t_0)$ since the columns of the Wronskian are all linearly independent (linearly independent solutions). The solution to the Cauchy problem is thus:

$$\mathbf{y}(t) = \mathbf{Y}(t) \cdot \mathbf{Y}^{-1}(t_0) \cdot \mathbf{y}_0$$

29.4 Non-homogeneous equation

Now that we have solved the homogeneous equation, we may generalise our results to non-homogeneous differential equations:

$$\mathbf{y}(t) = \mathbf{Y}(t) \cdot \mathbf{C} + \mathbf{y}_{PI}(t)$$

which subject to the initial condition $\mathbf{y}(t_0) = \mathbf{y}_0$ yields:

$$\mathbf{C} = \mathbf{Y}^{-1}(t_0) \cdot [\mathbf{y}_0 - \mathbf{y}_{PI}(t_0)]$$

So how can we find the particular integral? It suffices to use the method of variation of constants, $C_i \rightarrow \psi(t)$:

$$\mathbf{y}(t) = \mathbf{y}_i(t)\psi_i(t) = \mathbf{Y}(t) \cdot \psi \quad (29.4.1)$$

We substitute this into the non homogeneous equation:

$$\dot{\mathbf{y}}(t) = \dot{\mathbf{Y}} \cdot \psi + \mathbf{Y} \cdot \dot{\psi} = A \cdot \mathbf{Y} \cdot \psi + \mathbf{f} \implies \mathbf{Y} \cdot \dot{\psi} = \mathbf{f}$$

since $\dot{\mathbf{Y}} = A \cdot \mathbf{Y}$. Assuming that $W(t) \neq 0$, then:

$$\dot{\psi} = \mathbf{Y}^{-1} \cdot \mathbf{f} \implies \psi = \mathbf{C} + \int_{t_0}^t \mathbf{Y}^{-1}(t') \cdot \mathbf{f}(t') dt'$$

so that:

$$\mathbf{y}(t) = \mathbf{Y}(t) \cdot \underbrace{\left[\mathbf{C} + \int_{t_0}^t \mathbf{Y}^{-1}(t') \cdot \mathbf{f}(t') dt' \right]}_{\text{CF}}$$

We substitute back the expression for \mathbf{C} and find:

$$\mathbf{y}(t) = \mathbf{Y}(t) \cdot \left[\mathbf{Y}^{-1}(t_0) \cdot \mathbf{y}_0 + \int_{t_0}^t \mathbf{Y}^{-1}(t) \cdot \mathbf{f}(t') dt' \right]$$

29.5 Higher Order ODEs

Consider the general form of a linear ODE:

$$y^{(n)} + p_{n-1}(t)y^{(n-1)} + \dots + p_1(t)y' + p_0(t)y = f(t) \quad (29.5.1)$$

and assume we have found a fundamental set of solutions $\{y_i(t)\}$ for the associated homogeneous differential equation. Then, by the principle of superposition, the complementary function is:

$$y_{CF}(t) = \sum_{i=1}^n c_i y_i(t)$$

We now *weaponize* the constants and find:

$$y_{PI}(t) = \sum_{i=1}^n \psi_i y_i \quad (29.5.2)$$

and assuming that $\sum_{i=1}^n \psi'_i y_i = 0$, differentiating yields:

$$y'_{PI}(t) = \sum_{i=1}^n \psi y'_i$$

In general, we will set:

$$\begin{cases} y_{PI}^{(k)}(t) = \sum_{i=1}^n \psi_i y_i^{(k)}, & k = 1, \dots, n-1 \\ \sum_{i=1}^n \psi'_i y_i^{(k)} = 0, & k = 0, \dots, n-2 \end{cases}$$

Finally, we evaluate the n th derivative as usual without making any special assumptions:

$$y_{PI}^{(n)}(t) = \sum_{i=1}^n (\psi_i y_i^{(n)} + \psi'_i y_i^{(n-1)})$$

We are now ready to substitute everything into (5.5):

$$\sum_{i=1}^n (\psi_i y_i^{(n)} + \psi'_i y_i^{(n-1)}) + p_{n-1}(t) \sum_{i=1}^n \psi_i y_i^{(n-1)} + \dots + p_1(t) \sum_{i=1}^n \psi_i y'_i + p_0(t) \sum_{i=1}^n \psi_i y_i = f(t)$$

which rearranging gives:

$$\sum_{i=1}^n \left(\psi_i \left[\sum_{j=0}^n p_j y_i^{(j)} \right] \right) + \sum_{i=1}^n \psi'_i y_i^{(n-1)} = f(t)$$

Note that since $\{y_i(t)\}$ are all solutions to the associated homogeneous equation, then $\sum_{j=0}^n p_j y_i^{(j)} = 0$, so that:

$$\sum_{i=1}^n \psi'_i y_i^{(n-1)} = f(t)$$

We therefore have the following system of equations:

$$\begin{cases} \sum_{i=1}^n \psi'_i y_i^{(k)} = 0, \quad k = 0, \dots, n-2 \\ \sum_{i=1}^n \psi'_i y_i^{(n-1)} = f(t) \end{cases}$$

To solve this system of equations, we will use Cramer's rule. As always, the Wronskian is:

$$W(t) = \begin{vmatrix} y_1 & y_2 & \dots & y_n \\ y'_1 & y'_2 & \dots & y'_n \\ \vdots & \vdots & & \vdots \\ y_1^{(n-1)} & y_2^{(n-1)} & \dots & y_n^{(n-1)} \end{vmatrix}$$

To use Cramer's rule, we must successively substitute each column of the Wronskian with $(0 \ 0 \ \dots \ f(t))^T$. We will thus denote:

$$W_i = \begin{vmatrix} y_1 & \dots & y_{i-1} & 0 & \dots & y_n \\ y'_1 & \dots & y'_{i-1} & 0 & \dots & y'_n \\ \vdots & & \vdots & & \vdots & \\ y_1^{(n-1)} & \dots & y_{i-1}^{(n)} & f(t) & \dots & y_n^{(n-1)} \end{vmatrix} = g(t) \begin{vmatrix} y_1 & \dots & y_{i-1} & 0 & \dots & y_n \\ y'_1 & \dots & y'_{i-1} & 0 & \dots & y'_n \\ \vdots & & \vdots & & \vdots & \\ y_1^{(n-1)} & \dots & y_{i-1}^{(n)} & 1 & \dots & y_n^{(n-1)} \end{vmatrix}$$

Cramer's rule finally gives:

$$\psi'_i = \frac{g(t)W_i(t)}{W(t)} \implies u_i = \int \frac{g(t)W_i(t)}{W(t)} dt$$

which substituting back into (5.6) gives:

$$y_{PI}(t) = \sum_{i=1}^n \left(y_i(t) \int \frac{g(t)W_i(t)}{W(t)} dt \right) \quad (29.5.3)$$

Systems of Linear Differential Equations

In chapter 5, we learned how to solve the Cauchy problem for any system of linear ODEs, provided we knew the fundamental system of solutions. We are now going to see, for the special case of constant coefficients, how to find this fundamental set.

30.1 Non-degenerate Eigenvalues

Consider the system of ODEs:

$$\dot{\mathbf{y}} = \mathbf{A} \cdot \mathbf{y} + \mathbf{f}(t) \quad (30.1.1)$$

To find the eigenvalues of the matrix \mathbf{A} , it suffices to solve the *eigenvalue equation*:

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0 \quad (30.1.2)$$

and then find the corresponding eigenvectors using the *eigenvector equation*:

$$(\mathbf{A} - \lambda \mathbf{I}) \cdot \mathbf{v}_i = 0. \quad (30.1.3)$$

Proposition 3.

If a matrix has non-degenerate eigenvalues λ_i , then the corresponding eigenvectors \mathbf{v}_i form a basis.

Proof. Consider (6.3), if $\{\mathbf{v}_i\}$ are linearly dependent, then it must be possible to write one as:

$$\mathbf{v}_n = \sum_j \alpha_j \mathbf{v}_j$$

Then, on one hand we find:

$$\mathbf{A} \cdot \mathbf{v}_n = \lambda_n \mathbf{v}_n = \lambda_n \sum_j \alpha_j \mathbf{v}_j$$

and on the other hand:

$$\mathbf{A} \cdot \mathbf{v}_n = \sum_j \alpha_j \mathbf{A} \cdot \mathbf{v}_j = \sum_j \lambda_j \alpha_j \mathbf{v}_j.$$

Equating the two gives:

$$\sum_j \alpha_j \mathbf{v}_j (\lambda_n - \lambda_j) = 0 \implies \alpha_j = 0$$

since by assumption, the eigenvectors are independent and the eigenvalues are non degenerate. ■

Following the algebraic approach of the Wronskians, we pack the eigenvectors and eigenvalues into matrices as:

$$\begin{aligned} \mathbf{R} &\equiv (\mathbf{v}_1 \dots \mathbf{v}_n), \quad \mathbf{L} \equiv \text{diag}\{\lambda_i\} \\ \implies \mathbf{A} \cdot \mathbf{R} &= \mathbf{R} \cdot \mathbf{L} \\ \implies \mathbf{A} &= \mathbf{R} \cdot \mathbf{L} \cdot \mathbf{R}^{-1} \end{aligned}$$

We may substitute this into the associated homogeneous equation to (6.1) and find:

$$\begin{aligned} \mathbf{y} &= \mathbf{R} \cdot \mathbf{L} \cdot \mathbf{R}^{-1} \cdot \mathbf{y} \\ \implies \frac{d}{dt}(\mathbf{R}^{-1} \cdot \mathbf{y}) &= \mathbf{L} \cdot \mathbf{R}^{-1} \cdot \mathbf{y} \end{aligned}$$

and hence we get:

$$\dot{\zeta} = \mathbf{L} \cdot \zeta \implies \zeta_i(t) = C_i e^{\lambda_i t}$$

We transform back into the original coordinates so that:

$$\mathbf{y} = \sum_i C_i \mathbf{v}_i e^{\lambda_i t} \tag{30.1.4}$$

For simplicity, define the following matrix:

$$E(t) \equiv \text{diage}^{\lambda_i t}$$

which recasts (6.4) as:

$$\mathbf{y} = \underbrace{\mathbf{R} \cdot E(t) \cdot \mathbf{C}}_{Y(t)}$$

Finally, let us set the Cauchy condition $\mathbf{y}(0) = \mathbf{y}_0$

$$\mathbf{y}_0 = \mathbf{R} \cdot \mathbf{C} \implies \mathbf{C} = \mathbf{R}^{-1} \cdot \mathbf{y}_0$$

and hence:

$$\mathbf{y} = \mathbf{R} \cdot E(t) \cdot \mathbf{R}^{-1} \cdot \mathbf{y}_0$$

30.2 Matrix exponentiation

Consider now the more general Cauchy problem (6.1), which can be solved, as always, via the method of variation of parameters:

$$\mathbf{y}(t) = \mathbf{Y}(t) \left[\mathbf{Y}^{-1}(t_0) \cdot \mathbf{y}_0 + \int_{t_0}^t \mathbf{Y}^{-1}(t') \cdot \mathbf{f}(t') dt' \right]$$

Here:

$$\begin{aligned} \mathbf{Y}^{-1}(t) &= \mathbf{E}^{-1}(t) \cdot \mathbf{R}^{-1} \implies \mathbf{Y}(t) \cdot \mathbf{Y}^{-1}(t') = \mathbf{R} \cdot \mathbf{E}(t) \cdot \mathbf{E}^{-1}(t) \cdot \mathbf{R}^{-1} = \mathbf{R} \cdot \mathbf{E}(t - t') \cdot \mathbf{R}^{-1} \\ \mathbf{Y}^{-1}(0) &= \mathbf{R}^{-1} \end{aligned}$$

so that:

$$\mathbf{y}(t) = \mathbf{Y}(t_0) \left[\mathbf{Y}^{-1}(t) \cdot \mathbf{y}_0 + \int_{t_0}^t \mathbf{R} \cdot \mathbf{E}(t - t') \cdot \mathbf{R}^{-1} \cdot \mathbf{f}(t') dt' \right]$$

We may now compare this with the one-dimensional solution using the integrating factor method, we realize that a new expression for matrix exponentiation can be found:

$$e^{\mathbf{A}t} = \mathbf{R} \cdot \mathbf{E}(t) \cdot \mathbf{R}^{-1} \quad (30.2.1)$$

30.3 Higher Order Constant Coefficient Equations

Consider the nth order linear ODE with constant coefficients:

$$y^{(n)} + a_{n-1}y^{(n-1)} + \dots + a_1\dot{y} + a_0y = f(t) \quad (30.3.1)$$

which can be reduced into a first order linear system of ODEs:

$$\begin{cases} \dot{p}_{n-1} = -a_{n-1}p_{n-1} - \dots - a_1p_1 - a_0y + f(t) \\ \dot{p}_{n-2} = p_{n-1} \\ \dots \\ \dot{p}_1 = p_2\dot{y} = p_1 \end{cases}$$

or in matrix form:

$$\frac{d}{dt} \begin{pmatrix} p_{n-1} \\ \vdots \\ p_1 \\ y \end{pmatrix} = \begin{pmatrix} -a_{n-1} & -a_{n-2} & \dots & -a_1 & -a_0 \\ 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} p_{n-1} \\ \vdots \\ p_1 \\ y \end{pmatrix} + \begin{pmatrix} f \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

It can then be shown that the eigenvalue equation and auxiliary equation are equivalent:

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0 \iff \lambda^n + a_{n-1}\lambda^{n-1} + \dots + \lambda a_1 + a_0 = 0 \quad (30.3.2)$$

If the matrix \mathbf{A} turns out to be diagonalisable, then the solution is:

$$y(t) = \sum_{i=1}^n C_i e^{\lambda_i t}$$

However, if the matrix is not diagonalisable, and the eigenvalues are thus degenerate and don't have distinct eigenvalues, then this approach will not work.

30.4 Triangulation

Theorem 4. (Schur's Triangulation Theorem)

For any matrix \mathbf{A} , there is a unitary transformation that converts it into triangular form:

$$\mathbf{U}^\dagger \cdot \mathbf{A} \cdot \mathbf{U} = \mathbf{T} = \begin{pmatrix} \lambda_1 & & \text{stuff} \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \quad (30.4.1)$$

Proof. Consider one eigenvector of our matrix:

$$\mathbf{A} \cdot \mathbf{v}_1 = \lambda_1 \mathbf{v}_1$$

and consider an orthonormal basis $\{\mathbf{w}_i\}$ with $\mathbf{w}_1 = \mathbf{v}_1$ and $\mathbf{w}_i^* \cdot \mathbf{w}_j = \delta_{ij}$. As a consequence of this orthonormality, the matrix:

$$\mathbf{R} = (\mathbf{v}_1 \ \mathbf{w}_2 \ \dots \ \mathbf{w}_n)$$

is unitary, that is, $\mathbf{R}^{-1} = \mathbf{R}^\dagger$. Therefore, it follows that:

$$\mathbf{R}^\dagger \cdot \mathbf{A} \cdot \mathbf{R} = \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_n \end{pmatrix} \cdot \mathbf{R} = (\lambda_1 \mathbf{v}_1 \ \mathbf{A} \cdot \mathbf{w}_2 \ \dots \ \mathbf{A} \cdot \mathbf{w}_n) = \begin{pmatrix} \lambda_1 \ \mathbf{v}_1^* \cdot \mathbf{A} \cdot \mathbf{w}_2 & \dots & \mathbf{v}_1^* \cdot \mathbf{A} \cdot \mathbf{w}_2 \\ 0 & & \\ \vdots & & \mathbf{A}_{n-1} \\ 0 & & \end{pmatrix}$$

where \mathbf{A}_{n-1} has elements $\mathbf{w}_i^* \cdot \mathbf{A} \cdot \mathbf{w}_j$. We can keep on repeating this to \mathbf{A}_{n-1} until we reach an upper triangular matrix as required. ■

Once we have found the unitary transformation made up of an eigenvector and an orthonormal basis, we can then find a solution to the system of ODEs:

$$\dot{\mathbf{y}} = \mathbf{A} \cdot \mathbf{y} = \mathbf{U} \cdot \mathbf{T} \cdot \mathbf{U}^\dagger \implies \frac{d}{dt}(\mathbf{U}^\dagger \cdot \mathbf{y}) = \mathbf{T} \cdot \underbrace{\mathbf{U}^\dagger \cdot \mathbf{y}}_{\zeta} \implies \dot{\zeta} = \mathbf{T} \cdot \zeta$$

where:

$$T = \begin{pmatrix} \lambda_1 & T_{12} & T_{13} & \dots & T_{1n} \\ 0 & \lambda_2 & T_{23} & \dots & T_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & \lambda_n \end{pmatrix}$$

This system can be solved component by component, since it is made up of first order ODEs. Once ζ has been found, simply revert back to y :

$$y = U \cdot \zeta$$

30.5 Jordan Form

Finally, let us consider a simpler method for solving ODEs with non-diagonalisable matrices.

Definition 6.1 Consider a degenerate eigenvalue λ_1 , called *generator*, that repeats m times, with at least one associated eigenvector v_1 . Then, $\{v_1 \dots v_k\}$ is called a *Jordan Chain* if they satisfy:

$$\begin{cases} A \cdot v_1 = \lambda_1 v_1 \\ A \cdot v_2 = \lambda_1 v_2 + v_1 \\ \dots \\ A \cdot v_k = \lambda_1 v_k + v_{k-1} \end{cases}$$

and are therefore linearly independent. We may do so for a set of eigenvalues $\{\lambda_i\}$ and find the Jordan form:

$$R^{-1} \cdot A \cdot R = J = \text{diag } J_i$$

where for each eigenvalue λ_1 corresponds a Jordan block:

$$J_i = \begin{pmatrix} \lambda_i & 1 & 0 & \dots & 0 \\ 0 & \lambda_i & 0 & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \dots & \lambda_i \end{pmatrix}$$

Theorem. (Jordan's Theorem)

For any matrix A, there is always a basis in \mathbb{C}^n consisting of Jordan chains.

As usual, we have:

$$\dot{y} = A \cdot y = R \cdot J \cdot R^{-1} \implies \dot{\zeta} = J \cdot \zeta$$

where $\zeta = R^{-1} \cdot y$.

For example, consider the Jordan block J_1 :

$$\left\{ \begin{array}{l} \dot{\zeta}_1 = \lambda_1 \zeta_1 + \zeta_2 \\ \dots \\ \dot{\zeta}_{k-1} = \lambda_1 \zeta_{k-1} + \zeta_k \\ \dot{\zeta}_k = \lambda_1 \zeta_k \end{array} \right. \implies \left\{ \begin{array}{l} \zeta_k = C_k e^{\lambda_1 t} \\ \zeta_{k-1} = (C_{k-1} + C_k t) e^{\lambda_1 t} \\ \dots \\ \zeta_1 = \left(C_1 + C_2 t + \dots + C_k \frac{t^{k-1}}{(k-1)!} \right) e^{\lambda_1 t} \end{array} \right.$$

Finally, reverting back to \mathbf{y} :

$$\mathbf{y}(t) = \zeta_1 \mathbf{v}_1 + \dots + \zeta_k \mathbf{v}_k \quad (30.5.1)$$

Series Solutions and special functions

31.1 Power Series

We summarise below a set of properties of power series:

1. The power series $\sum_{n=0}^{\infty} a_n(x - x_0)^n$ converges at a point x if the following limit exists:

$$\lim_{n \rightarrow \infty} \sum_{n=0}^{\infty} a_n(x - x_0)^n$$

and converges absolutely at a point x if the associated power series:

$$\sum_{n=0}^{\infty} |a_n(x - x_0)|^n$$

converges.

2. If a power series converges absolutely, it converges. The converse isn't always true.
3. If $a_n \neq 0$ and for a fixed x :

$$L = \lim_{n \rightarrow \infty} |x - x_0| \left| \frac{a_{n+1}}{a_n} \right|$$

then if $L < 1$, the series converges, if $L > 1$, the series diverges, and $L = 1$ is inconclusive.

4. For a power series, there exists $0 \leq \rho \leq \infty$ called *radius of convergence* such that it will converge for $|x - x_0| < \rho$ and diverge for $|x - x_0| > \rho$.
5. Suppose $\sum_{n=0}^{\infty} a_n(x - x_0)^n$ converges to $f(x)$. Then f is continuous and is infinitely differentiable over the interval of convergence $|x - x_0| < \rho$.
6. The value of a_n is then given by:

$$a_n = \frac{f^{(n)}(x_0)}{n!}$$

and the series is then called the *Taylor series*:

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!} (x - x_0)^n$$

A function with Taylor series with non-zero radius of convergence is said to be analytic.

7. If $\sum_{n=0}^{\infty} a_n(x - x_0)^n = \sum_{n=0}^{\infty} b_n(x - x_0)^n$, then $a_n = b_n$ for all n.

31.2 Series Solutions near ordinary points

Let us consider the second order linear homogeneous ODE:

$$P(x)y'' + Q(x)y' + R(x)y = 0 \quad (31.2.1)$$

Definition 7.1 A point x_0 such that $P(x_0) \neq 0$ is an *ordinary point*.

A point x_0 such that $P(x_0) = 0$ is an *singular point*.

Since P is continuous, we may choose an interval containing x_0 , and divide (6.1) through by $P(x)$ obtaining:

$$y'' + p(x)y' + q(x)y = 0$$

where $p(x) = \frac{Q(x)}{P(x)}$, $q(x) = \frac{R(x)}{P(x)}$. We look for solutions of the form:

$$y = \sum_{n=0}^{\infty} a_n(x - x_0)^n$$

with radius of convergence $\rho > 0$. One can then substitute this expression into the ODE and use the aforementioned properties of power series to deduce the coefficients.

Theorem 5.

If x_0 is an ordinary point of:

$$P(x)y'' + Q(x)y' + R(x)y = 0,$$

then the general solution is:

$$y = \sum_{n=0}^{\infty} a_n(x - x_0)^n = a_0 y_1 + a_1 y_2$$

Further, the solutions y_1 and y_2 form a fundamental set of solutions, and the radius of convergence for each of them is greater than or equal to the minimum of the radii of convergence of the series for $p = \frac{Q(x)}{P(x)}$ and $q = \frac{R(x)}{P(x)}$:

$$\rho_{y_1} \geq \min\{\rho_p, \rho_q\}.$$

31.3 Airy's Equation

For example, let us solve *Airy's equation*

$$y'' - xy = 0, \quad -\infty < x < \infty.$$

We note that $P(x) = 1, Q(x) = 0, R(x) = -x$, hence any point is ordinary. We assume that:

$$y = \sum_{n=0}^{\infty} a_n(x)^n$$

converges in some interval $|x| < \rho$. Then:

$$\sum_{n=0}^{\infty} (n+2)(n+1)a_{n+2}x^n - \sum_{n=0}^{\infty} a_n(x)^{n+1} = 0$$

and using a shift of index we may rewrite:

$$2a_2 + \sum_{n=1}^{\infty} (n+2)(n+1)a_{n+2}x^n - \sum_{n=1}^{\infty} a_{n-1}(x)^n = 0.$$

This is only possible for all x if the coefficients of like powers of x cancel each other out:

$$(n+2)(n+1)a_{n+2} - a_{n-1} = 0, \quad \text{for } n = 1, 2, 3\dots$$

This is a second order recurrence relation, so the coefficients will be determined in steps of 3. Note that:

$$a_2 = 0 \implies a_{3n+2} = 0$$

Furthermore:

$$a_3 = \frac{a_0}{2 \cdot 3}, \quad a_6 = \frac{a_3}{5 \cdot 6} = \frac{a_0}{2 \cdot 3 \cdot 5 \cdot 6}, \quad a_9 = \frac{a_6}{8 \cdot 9} = \frac{a_0}{2 \cdot 3 \cdot 5 \cdot 6 \cdot 8 \cdot 9}$$

which suggests (as can be proven by induction):

$$a_{3n} = \frac{a_0}{2 \cdot 3 \cdot 5 \cdot 6 \dots (3n-1)(3n)}$$

Finally, we also find:

$$a_4 = \frac{a_1}{3 \cdot 4}, \quad a_7 = \frac{a_4}{6 \cdot 7} = \frac{a_1}{3 \cdot 4 \cdot 6 \cdot 7}, \quad a_{10} = \frac{a_7}{9 \cdot 10} = \frac{a_0}{3 \cdot 4 \cdot 6 \cdot 7 \cdot 9 \cdot 10}$$

which suggests:

$$a_{3n+1} = \frac{a_1}{3 \cdot 4 \cdot 6 \cdot 7 \dots (3n)(3n+1)}$$

Finally, we arrive at the general solution to Airy's equation:

$$y(x) = a_0 \left(1 + \sum_{n=1}^{\infty} \frac{x^{3n}}{2 \cdot 3 \cdot 5 \cdot 6 \dots (3n-1)(3n)} \right) + a_1 \left(x + \sum_{n=1}^{\infty} \frac{x^{3n+1}}{3 \cdot 4 \cdot 6 \cdot 7 \dots (3n)(3n+1)} \right) \quad (31.3.1)$$

Also, notice that the first sum satisfies $y(0) = 0, y'(0) = 0$ and the second sum satisfies $y(0) = 0, y'(0) = 1$, which implies that $W(0) = 1 \neq 0$, and hence (7.2) truly is the general solution.

Sturm-Louiville theory

32

Distributions

33.1 Introducing the Dirac delta

Suppose we have a random variable x with a gaussian probability distribution:

$$\rho(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2} \quad (33.1.1)$$

Note that this being a probability density function implies that it is properly normalized, and thus satisfies:

$$\int_{\mathbb{R}} \rho(x) dx = 1 \quad (33.1.2)$$

We ask what would happen if we let $\sigma \rightarrow 0$? Clearly we must have that for $x \neq 0$, $\lim_{\sigma \rightarrow 0} \rho(x \neq 0) = 0$ due to the exponential suppression overpowering the $1/\sigma$ factor in the front. However, we must still have that:

$$\lim_{\sigma \rightarrow 0} \int_{\mathbb{R}} \rho(x) dx = 1 \quad (33.1.3)$$

so clearly the value of this function at 0 must be very peculiar.

Definition (Dirac delta function) We define the Dirac delta function to be the limit of a normalized gaussian in the zero standard deviation limit:

$$\delta(x) \equiv \lim_{\sigma \rightarrow 0} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2} \quad (33.1.4)$$

which satisfies:

$$\begin{cases} \delta(x) = 0, & x \neq 0 \\ \int_{\mathbb{R}} \delta(x) dx = 1 \end{cases} \quad (33.1.5)$$

Statistically, the Dirac delta represents the probability distribution of a random variable which is certainly equal to zero.

Let's now consider:

$$\mathbb{P}(X \leq x) = \int_{-\infty}^x \rho(x') dx' \quad (33.1.6)$$

which is the probability of the variable X having a value smaller than x . As long as $\sigma > 0$ we find that:

$$\frac{d\mathbb{P}(X \leq x)}{dx} = \rho(x') \quad (33.1.7)$$

Should we expect this to hold for $\sigma \rightarrow 0$? In this limit we have that:

$$\mathbb{P}(X \leq x) \rightarrow \Theta(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \quad (33.1.8)$$

since the probability of X being smaller than one should be zero. Extending (33.1.7) in the $\sigma \rightarrow 0$ case gives:

$$\frac{d\Theta(x)}{dx} = \delta(x) \quad (33.1.9)$$

We would like to make this reasoning more rigorous.

To restate our point, consider for example a wire strung between two walls at $x = \pm L$, taught with tension T under a weight W hung at $x = 0$.

We should expect there to be no curvature in the wire away from the mass:

$$\frac{d^2y}{dx^2} = 0, \quad x \neq 0 \quad (33.1.10)$$

Also, $y(\pm L) = 0$ and at $x = a$ the wire's profile must be continuous (or else it would break). Newton's second law gives us:

$$2T \sin \theta \approx 2T \tan \theta = 2T \left[\frac{dy}{dx} \right]_{x=-L}^{x=L} = W \quad (33.1.11)$$

which can be solved:

$$y(x) = \begin{cases} -W(L+x)/2T, & -L < x < 0 \\ -W(L-x)/2T, & 0 < x < L \end{cases} \quad (33.1.12)$$

Note that this solution satisfies:

$$T \frac{d^2y}{dx^2} = W \delta(x) \quad (33.1.13)$$

Theorem (Dirac delta) For an object $\delta(x)$ satisfying (33.1.5), letting $f(x) \in L^1(\mathbb{R})$ then:

$$\int_R f(x)\delta(x)dx = f(0) \quad (33.1.14)$$

Proof. Let $\epsilon > 0$ be small. Then we have that:

$$\int_{\mathbb{R}} f(x)\delta(x)dx = \int_{-\epsilon}^{\epsilon} f(x)\delta(x)dx \quad (33.1.15)$$

since $\delta(x) = 0$ for $x \neq 0$. Next:

$$\int_{\mathbb{R}} f(x)\delta(x)dx = \int_{-\epsilon}^{\epsilon} (f(x) - f(0) + f(0))\delta(x)dx \quad (33.1.16)$$

$$= \int_{-\epsilon}^{\epsilon} (f(x) - f(0))\delta(x)dx + f(0) \quad (33.1.17)$$

Since $f(x)$ is continuous we must have that for any $\varepsilon' > 0$ there is an $\epsilon > 0$ such that:

$$|f(x) - f(0)| < \varepsilon', \quad \forall |x| < \epsilon \quad (33.1.18)$$

so that:

$$\lim_{\epsilon \rightarrow \infty} \int_{-\epsilon}^{\epsilon} (f(x) - f(0))\delta(x)dx = 0 \quad (33.1.19)$$

and thus

$$\int_{\mathbb{R}} f(x)\delta(x)dx = f(0) \quad (33.1.20)$$

as desired. ■

33.2 Rigorous treatment

Definition (Test function) A function $\phi : \mathbb{R} \rightarrow \mathbb{R}$ is a test function if:

- (i) $\phi(x) \in C^\infty$ (smoothness)
- (ii) $\phi(x) = 0$ outside of the disc $D(X) = \{|x| \leq X\}$ (compact support)

This definition ensures that $\phi(x) \rightarrow 0$ at $\pm\infty$ and also ensures that $\phi^{(n)}(x)$ is also a test function. These functions are called test functions because they are used to test the action of distributions (which we have yet to define) on them.

One famous example of a test function can be constructed from:

$$\Phi(x) = \begin{cases} 0, & x \leq 0 \\ e^{-1/x}, & x > 0 \end{cases} \quad (33.2.1)$$

Note that this function is infinitely differentiable (and they all vanish at $x = 0$ since its n th derivative will be to leading order $o(x^n/e^x)$). Unfortunately $\Phi(x)$ is not yet a test function since it does not vanish at $+\infty$. We can solve this by defining:

$$\phi(x) = \Phi(x)\Phi(1-x) \quad (33.2.2)$$

which does indeed satisfy the requirements of a test function.

Theorem (Unique definition of distribution) Let $f_{1,2} : \mathbb{R} \rightarrow \mathbb{R}$ be continuous functions. If $\langle f_1, \phi \rangle = \langle f_2, \phi \rangle$ for all test functions ϕ then $f_1(x) = f_2(x)$.

Proof. We begin by proving that if:

$$\langle f, \phi \rangle = \int_{\mathbb{R}} f(x)\phi(x) dx = 0 \quad (33.2.3)$$

for all test functions then $f(x) = 0$. Suppose that $f(a) > 0$ at some point $x = a$. Due to the continuity of f there exists $\delta > 0$ such that $f(x) > 0$ for all $x \in (a - \delta, a + \delta)$. We can find a test function $\phi(x)$ which vanishes outside $(a - \delta, a + \delta)$ and is non-zero inside this interval, implying that:

$$\langle f, \phi \rangle = \int_{a-\delta}^{a+\delta} f(x)\phi(x) dx > 0 \quad (33.2.4)$$

This however is a contradiction, so we cannot have that $f(a) > 0$ at some $x = a$. ■

Definition (Convergence of test function sequences) The sequence $\{\phi_n\}$ of test functions converges to zero if:

- (i) $\phi_n(x) = 0$ for all n and x outside some interval $I \subset \mathbb{R}$.
- (ii) for all k , $\phi_n^{(k)}$ converges uniformly to ∞ as $n \rightarrow \infty$.

33.3 Distributions

Definition (Distribution) A distribution (or generalised function) f is a continuous functional mapping from the set of test functions \mathcal{D} to \mathbb{R} via the action:

$$\phi \rightarrow \langle f, \phi \rangle \in \mathbb{R} \quad (33.3.1)$$

It must satisfy:

- (i) Continuous: if $\phi_n \rightarrow 0$ then $\langle f, \phi_n \rangle \rightarrow 0$.
- (ii) Linear

An example of a distribution is the Heaviside distribution \mathcal{T} generated from the Heaviside function $\Theta(x)$:

$$\langle \mathcal{T}, \phi \rangle = \int_0^{\infty} \phi(x) dx \quad (33.3.2)$$

Clearly this is linear. It is continuous since if $\phi_n(x) \rightarrow 0$ then $\langle \mathcal{T}, \phi \rangle \rightarrow 0$ too.

Laplace transform methods

34.1 Basic definition and properties of the Laplace transform

The Laplace transforms is a very important type of integral transform which is a precursor to its cousin, the Fourier transform. It, like its cousin, can be used to solve both ordinary and partial differential equations with more immediacy.

Definition (Laplace transform) The Laplace transform maps a function $f(t) \in L^p(\mathbb{R}^+)$ to another function $F(s) = \mathcal{L}\{f(t)\}$ defined by:

$$F(s) = \mathcal{L}\{f(t)\} \equiv \int_0^\infty e^{-st} f(t) dt \quad (34.1.1)$$

It is important to note that the Laplace transform is linear due to the linearity of integration so that:

$$\mathcal{L}\{\alpha f(t) + \beta g(t)\} = \alpha \mathcal{L}\{f(t)\} + \beta \mathcal{L}\{g(t)\} \quad (34.1.2)$$

Example. Compute the Laplace transform of the Heaviside function:

$$\Theta(t) = \begin{cases} 1, & t \geq 0 \\ 0, & t \leq 0 \end{cases} \quad (34.1.3)$$

We find that:

$$\mathcal{L}\{\Theta(t)\} = \int_0^\infty e^{-st} dt = -\frac{1}{s} \lim_{\tau \rightarrow \infty} (e^{-\tau s} - 1) = \frac{1}{s}, \quad s > 0 \quad (34.1.4)$$

where we must assume that $s > 0$ for the Laplace transform to be well defined. By the linearity of \mathcal{L} we may write:

$$\mathcal{L}\{c\Theta(t)\} = \frac{c}{s} \quad (34.1.5)$$

Note also that for any function $f(t)$ integrable on $[0, \infty)$:

$$\mathcal{L}\{f(t)\Theta(t)\} = \mathcal{L}\{f(t)\} \quad (34.1.6)$$

so from (34.1.5) we have found that:

$$\mathcal{L}\{c\} = \frac{c}{s} \quad (34.1.7)$$



Note that when performing the Laplace transform of some function we lose all information about its behaviour for $t < 0$, we say that the Laplace transform is not unitary. Consequently, one should not expect to be able to invert the Laplace transform and get back the initial function in its original domain since we don't know how it behaves for $t < 0$. There is an ambiguity in choosing the function's behaviour if we want to define it over \mathbb{R} . We can solve this problem by using the Heaviside function. Namely:

$$\mathcal{L}\{f(t)\Theta(t)\} = F(s) \implies \mathcal{L}\{F(s)\}^{-1} = f(t)\Theta(t) \quad (34.1.8)$$

We have no ambiguity here since we simply assume that $f(t) = 0, t < 0$.

Example. Compute the Laplace transform of $f(t) = t$.

We need to evaluate the following integral:

$$\mathcal{L}\{t\} = \int_0^\infty te^{-st} dt \quad (34.1.9)$$

which can be solved using Feynman's trick. Let:

$$I(s) = \int_0^\infty te^{-st} dt \quad (34.1.10)$$

Then:

$$I(s) = -\frac{d}{ds} \left(\int_0^\infty e^{-st} dt \right) = -\frac{d}{ds} \left(\frac{1}{s} \right) = \frac{1}{s^2}, \quad s > 0 \quad (34.1.11)$$

so we have that:

$$\mathcal{L}\{t\} = \mathcal{L}\{t\Theta(t)\} = \frac{1}{s^2}, \quad s > 0 \quad (34.1.12)$$



Example. Compute the Laplace transform of $f(t) = \sin(kt)$ and $g(t) = \cos(kt)$. It is useful to first evaluate $\mathcal{L}\{e^{kt}\}$. This is a trivial integral:

$$\mathcal{L}\{e^{kt}\} = \mathcal{L}\{e^{kt}\Theta(t)\} = \int_0^\infty e^{-(s-k)t} dt = \frac{1}{s-k}, \quad s > k \quad (34.1.13)$$

It follows that:

$$\mathcal{L}\{e^{ikt}\} = \mathcal{L}\{\cos kt\} + i\mathcal{L}\{\sin kt\} = \frac{s}{s^2 + k^2} + i\frac{k}{s^2 + k^2}, \quad s > 0 \quad (34.1.14)$$

implying that:

$$\mathcal{L}\{\cos kt\} = \frac{s}{s^2 + k^2}, \quad \mathcal{L}\{\sin kt\} = \frac{k}{s^2 + k^2}, \quad s > 0 \quad (34.1.15)$$

◀

Proposition (Properties of the Laplace transform) Given a function $f(t) \in L^p(\mathbb{R}^+)$ with Laplace transform:

$$F(s) = \mathcal{L}\{f(t)\} \quad (34.1.16)$$

we have that:

$$\mathcal{L}\{e^{kt} f(t)\} = F(s - k) \quad (34.1.17)$$

and

$$\mathcal{L}\{t^n f(t)\} = (-1)^n \frac{d^n F(s)}{ds^n} \quad (34.1.18)$$

Proof. Firstly:

$$\mathcal{L}\{e^{kt} f(t)\} = \int_0^\infty e^{-(s-k)t} f(t) dt = F(s - k) \quad (34.1.19)$$

Secondly:

$$\mathcal{L}\{t^n f(t)\} = \int_0^\infty t^n f(t) e^{-st} dt = (-1)^n \frac{d^n}{ds^n} \left(\int_0^\infty f(t) e^{-st} dt \right) = (-1)^n \frac{d^n F(s)}{ds^n} \quad (34.1.20)$$

■

Theorem (Laplace transform of a derivative)

Let $f(t)$ be a n -differentiable function at 0. Then if $\mathcal{L}\{f(t)\} = F(s)$ then:

$$\mathcal{L}\{f^{(n)}(t)\} = s^n F(s) - \sum_i^{n-1} s^{n-i} f^{(i)}(0) \quad (34.1.21)$$

Proof. We proceed by induction. We have that:

$$\mathcal{L}\{\dot{f}(t)\} = \int_0^\infty \dot{f}(t) e^{-st} dt = \int_0^\infty \frac{d}{dt} (f(t) e^{-st}) dt + \int_0^\infty s f(t) e^{-st} dt \quad (34.1.22)$$

$$= sF(s) + [f(t)e^{-st}]_0^\infty \quad (34.1.23)$$

$$= sF(s) - f(0) \quad (34.1.24)$$

assuming that $f(t)$ is dominated by the exponentially decaying e^{-st} as $t \rightarrow \infty$. Let us now suppose that (34.1.21) is true up to n . Then:

Proposition (Second shift theorem) We have that:

$$\mathcal{L}\{f(t-a)\Theta(t-a)\} = e^{-sa}F(s) \quad (34.1.25)$$

We have that:

$$\mathcal{L}\{f(t-a)\Theta(t-a)\} = \int_0^\infty \Theta(t-a)e^{-st}f(t-a)dt \quad (34.1.26)$$

$$= \int_a^\infty e^{-st}f(t-a)dt = e^{-sa} \int_0^\infty e^{-st'}f(t')dt' = e^{-sa}F(s) \quad (34.1.27)$$

34.2 Solving ODEs with Laplace transforms

Consider the general linear inhomogeneous second order ODE with constant coefficients:

$$a\ddot{x} + b\dot{x} + cx = f(t), \quad x(0) = x_0, \dot{x}(0) = 0 \quad (34.2.1)$$

where $f(t)$ has Laplace transform $F(s)$. We can take the Laplace transform of this equation:

$$a(s^2\mathcal{L}\{x(t)\} - sx_0 - 0) + b(s\mathcal{L}\{x(t)\} - x_0) + c\mathcal{L}\{x(t)\} = F(s) \quad (34.2.2)$$

$$\implies (as^2 + bs + c)\mathcal{L}\{x(t)\} - asx_0 - bsx_0 = F(s) \quad (34.2.3)$$

$$\implies \mathcal{L}\{x(t)\} = \frac{F(s) + asx_0 + bsx_0}{as^2 + bs + c} \quad (34.2.4)$$

It is interesting to note that the characteristic polynomial of the ODE popped up in the denominator of this Laplace transform! For the homogeneous case where $f(t) = 0 \implies F(s) = 0$, assuming that the characteristic polynomial has roots at $\lambda_{1,2}$ then we find that:

$$\mathcal{L}\{x(t)\} = \frac{asx_0 + bsx_0}{(s - \lambda_1)(s - \lambda_2)} = \frac{A}{(s - \lambda_1)} + \frac{B}{(s - \lambda_2)} \quad (34.2.5)$$

where

$$A = \frac{a\lambda_1 - b}{\lambda_1 - \lambda_2}x_0, \quad B = \frac{b - a\lambda_2}{\lambda_1 - \lambda_2}x_0 \quad (34.2.6)$$

We can invert the Laplace transform and find that:

$$x(t) = Ae^{\lambda_1 t} + Be^{\lambda_2 t} \quad (34.2.7)$$

as expected. The importance of Laplace transforms is now clear: it is a very useful tool in solving differential equations.

Phase plane analysis

35

First order PDEs

36

Second order PDEs: a general overview

37

Parabolic PDEs: Diffusion

38

Hyperbolic PDEs: Waves

39

Elliptic PDEs: Electrostatics

40

Part V

Linear Algebra

Vector spaces

41.1 Definitions

We begin by defining a fundamental mathematical concept used in several areas of physics (most notably Quantum Mechanics), the **Vector Space**.

Classically speaking, vectors are defined as objects with both a magnitude and direction. However, as we will see soon this definition is very limited, and breaking beyond the barrier of arrows with lengths and directions will enable us to create a broader mathematical structure.

Definition 34.1 (*Vector space axioms*)

A linear space V over a field \mathbb{K} is a collection of **vectors** \mathbf{v} over which two binary operations $+$, \cdot are defined, such that $\forall \mathbf{u}, \mathbf{v}, \mathbf{z} \in V$ and $\forall \alpha_1, \alpha_2 \in \mathbb{K}$ the following are satisfied:

- (VS1) Closure under addition: $\mathbf{u} + \mathbf{v} \in V$
- (VS2) Closure under scalar multiplication: $\alpha_1 \mathbf{u} \in V$
- (VS3) Commutativity of addition: $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$
- (VS4) Associativity of addition: $\mathbf{u} + \mathbf{v}$
- (VS5) Associativity of addition: $\mathbf{u} + (\mathbf{v} + \mathbf{z}) = (\mathbf{u} + \mathbf{v}) + \mathbf{z}$
- (VS6) Associativity of scalar multiplication: $\alpha_1(\alpha_2 \mathbf{u}) = \alpha_1 \alpha_2 \mathbf{u}$
- (VS7) Right-distributivity: $(\alpha_1 + \alpha_2) \mathbf{u} = \alpha_1 \mathbf{u} + \alpha_2 \mathbf{u}$
- (VS8) Left-distributivity: $\alpha_1(\mathbf{u} + \mathbf{v}) = \alpha_1 \mathbf{u} + \alpha_1 \mathbf{v}$
- (VS9) Existence of **zero vector**: $\exists \mathbf{0} \in V$ such that $\mathbf{u} + \mathbf{0} = \mathbf{u}$
- (VS10) Existence of inverse under addition: $\exists (-\mathbf{u}) \in V$ such that $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$

Definition 34.2 (*Vector subspace*)

A vector space W is said to be a vector subspace of a vector space V if $W \subseteq V$, and is a proper subspace if W is neither the zero subspace $\{\mathbf{0}\}$ nor V .

So, if a vector space satisfying VS1-VS10 is a subset of some other vector space, then it is a vector subspace. Luckily, given a subset of V , one does not necessarily have to prove that all the vector space axioms hold, since some of them hold for all subsets of V . Indeed,

it turns out that only VS1 and VS2 do not necessarily hold for a subset of a vector space. All the others do.

Proposition 34.3 (*Criteria of vector subspaces*)

A subset $W \subseteq V$ is said to be a **vector subspace** of V over \mathbb{K} iff:

- (S1) Closure under addition: $\forall \mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_1 + \mathbf{w}_2 \in W$
- (S2) Closure under multiplication: $\forall \alpha \in \mathbb{K}, \forall \mathbf{w} \in W, \alpha \mathbf{w} \in W$
- (S3) Identity inclusion: $\mathbf{0}_V \in W$, where $\mathbf{0}_V$ is the zero of V , such that $\mathbf{0}_V + \mathbf{w} = \mathbf{w}, \forall \mathbf{w} \in W$.

Proof. We proceed by showing that all the vector space axioms hold for W :

- (VS1) is equivalent to S1
- (VS2) is equivalent to S2
- (VS3-VS8) $\mathbf{w}_1, \mathbf{w}_2 \in W \implies \mathbf{w}_1, \mathbf{w}_2 \in V$. Hence, since VS3-VS8 hold for all vectors of V , they must necessarily hold for all vectors of W .
- (VS9) is equivalent to S3
- (VS10) implication of S2 with α being the negative identity of \mathbb{K} .

■

Definition 34.4 (*Span*)

The span of a set of vectors $\{\mathbf{v}_1 \dots \mathbf{v}_k\}$ is defined as the set of all their linear combinations:

$$\text{Span}(\mathbf{v}_1 \dots \mathbf{v}_k) \equiv \left\{ \sum_{i=1}^k \alpha_i \mathbf{v}_i : \forall \alpha_i \in \mathbb{K} \right\} \quad (41.1.1)$$

Definition 34.5 (*Linear independence*)

Let V be a vector space over \mathbb{K} and let $\alpha_1 \dots \alpha_k \in \mathbb{K}$. Then we say that the set of vectors $\{\mathbf{v}_1 \dots \mathbf{v}_k\}$ are linearly independent iff:

$$\sum_{i=1}^k \alpha_i \mathbf{v}_i = \mathbf{0} \implies \alpha_i = 0, \forall 1 \leq i \leq k \quad (41.1.2)$$

Otherwise, they are said to be linearly dependent.

Firstly note that by this definition (that uses *otherwise*) a set of vectors is either linearly dependent or linearly independent, it cannot be both or neither.

Also note that it suffices for only one coefficient of a set of vectors to not be zero for linear dependence to be satisfied. Linear independence occurs only when all coefficients α_i must be zero.

An immediate result is the following:

Proposition 34.5 (Linear dependence of sets containing $\mathbf{0}$)

Any set of vectors $\{\mathbf{0}, \mathbf{v}_1, \dots, \mathbf{v}_k\} \subseteq V$ is linearly dependent.

Proof. The set $\{\mathbf{0}, \mathbf{v}_1, \dots, \mathbf{v}_k\} \subseteq V$ cannot be linearly independent, since for $\alpha \neq 0$ we may write:

$$\alpha \cdot \mathbf{0} + \sum_{i=1}^k 0 \cdot \mathbf{v}_i = \mathbf{0} \quad (41.1.3)$$

Therefore, the vectors must be linearly dependent. ■

41.2 Basis and dimensions

Definition 34.6 (Basis)

A set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\} \subseteq V$ is said to be a basis of V iff:

- (B1) they are linearly independent
- (B2) they generate V : $\text{Span}(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k) = V$.

Then, $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ are said to be **basis vectors**.

So, given a basis $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$, it is always possible to write any vector in V as a linear combination of these basis vectors.

Proposition 34.7 (Uniqueness of linear combination)

Let $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ be a basis of a vector space V . Then, any vector $\mathbf{v} \in V$ can be expressed as:

$$\mathbf{v} = \sum_{i=1}^k \alpha_i \mathbf{v}_i \quad (41.2.1)$$

where α_i are uniquely determined.

Proof. Suppose that \mathbf{v} can be expressed as two different linear combinations:

$$\mathbf{v} = \sum_{i=1}^k \alpha_i \mathbf{v}_i = \sum_{i=1}^k \alpha'_i \mathbf{v}_i \quad (41.2.2)$$

Then:

$$\sum_{i=1}^k (\alpha_i - \alpha'_i) \mathbf{v}_i = \mathbf{0} \quad (41.2.3)$$

However, since \mathbf{v}_i are linearly independent by (B1), this implies that $\alpha_i = \alpha'_i$, which is a contradiction. \blacksquare

Theorem 34.8 (Steinitz Exchange theorem)

Let $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ be a basis of V , and let $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_l\} \subsetneq V$. If $l > k$, then $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_l$ are linearly dependent.

Proof. If $\mathbf{w}_1 = \mathbf{0}$, then by Proposition 34.5 $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_l$ are linearly dependent.

Suppose $\mathbf{w}_1 \neq \mathbf{0}$. Then, we may write:

$$\mathbf{w}_1 = \sum_{i=1}^k \alpha_i \mathbf{v}_i \iff \mathbf{v}_1 = \frac{1}{\alpha_1} (\mathbf{w}_1 - \sum_{i=2}^k \alpha_i \mathbf{v}_i) \quad (41.2.4)$$

where we assume without loss of generality that $\exists \alpha_1 \neq 0$, since otherwise the sets would be linearly dependent as desired. Hence:

$$\text{Span}\left(\left(\mathbf{w}_1 - \sum_{i=2}^k \alpha_i \mathbf{v}_i\right), \mathbf{v}_2, \dots, \mathbf{v}_k\right) = V \quad (41.2.5)$$

where we omit $\frac{1}{\alpha_1}$ since it is only a constant and will be lost when writing out the linear combination. Note however that $\sum_{i=2}^k \alpha_i \mathbf{v}_i$ has already been included in the other vectors in the span, and can therefore be ignored:

$$\text{Span}(\mathbf{w}_1, \mathbf{v}_2, \dots, \mathbf{v}_k) = V \quad (41.2.6)$$

We can repeat this process by replacing \mathbf{v}_j with:

$$\frac{1}{\alpha'_j} \left(\mathbf{w}_j - \sum_{i=1}^{j-1} \alpha'_i \mathbf{w}_i - \sum_{i=j+1}^k \alpha'_i \mathbf{v}_i \right), \forall 1 < j \leq l \quad (41.2.7)$$

so that, by similar logic to before:

$$\text{Span}(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{j-1}, \mathbf{w}_j - \sum_{i=1}^{j-1} \alpha'_i \mathbf{w}_i - \sum_{i=j+1}^k \alpha'_i \mathbf{v}_i, \mathbf{v}_{j+1}, \dots, \mathbf{v}_k) = V \quad (41.2.8)$$

Again, all the vectors in $\sum_{i=1}^{j-1} \alpha'_i \mathbf{w}_i$ and $\sum_{i=j+1}^k \alpha'_i \mathbf{v}_i$ have already been included in the Span, and can be neglected. Hence, we get:

$$\text{Span}(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_j, \mathbf{v}_{j+1}, \dots, \mathbf{v}_l) \quad (41.2.9)$$

Now since $l \geq k$, the end result of reiterating this algorithm will be:

$$\text{Span}(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k) = V \quad (41.2.10)$$

with $l - k$ remaining \mathbf{w}_i vectors. This means that the vectors which were left out can be expressed as a linear combination of $\mathbf{w}_1 \dots \mathbf{w}_k$. For example:

$$\mathbf{w}_{l-k} = \sum_i^k \beta_i \mathbf{w}_i \implies \mathbf{w}_{l-k} - \sum_i^k \beta_i \mathbf{w}_i = \mathbf{0} \quad (41.2.11)$$

implying linear dependence. ■

The contrapositive of the Exchange lemma also provides an interesting result which we shall use soon.

Proposition 34.9 (*Contrapositive of the exchange theorem*)

Let $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ be a basis of V , and let $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_l\} \subsetneq V$. If $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_l$ are linearly independent, then $l \leq k$.

Definition 34.10 (*Dimension of a finite dimensional vector space*)

The dimension of a finite dimensional vector space is the number of vectors in its basis. So if V has a basis of cardinality n , then the dimension over a field \mathbb{K} is :

$$\dim_{\mathbb{K}}(V) \equiv n \quad (41.2.12)$$

It may not be immediately clear that the dimension of a vector space is well-defined. How can we know that all the bases of a vector space contain the same number of vectors?

Theorem 34.11 (*Well-definedness of vector space dimension*)

Any two bases of a finite dimensional vector space must contain the same number of basis vectors.

Proof. Suppose we have two bases, $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m\}$ and $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$, each of cardinality m and n respectively. Consequently, by B1, they must be both linearly independent. Using Proposition 34.9 then, we have that $m \geq k$ and $k \geq m$, implying that $k = m$ as desired. ■

It is interesting to note that the dimension of a vector space can depend on the field we define it on. For example, the vector space of all 3×3 matrices over \mathbb{R} has dimension 9, whereas over \mathbb{C} it has dimension 18.

In general, we will omit inserting the field when it is clear from the context.

Proposition 34.11 (Properties of finite dimensional spaces)

The following properties are satisfied by any finite dimensional vector space V :

- (D1) V has a basis
- (D2) every linearly independent subset of V can be expanded to form a basis
- (D3) If $n = \dim(V)$, then any linearly independent subset $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\} \subseteq V$ is a basis of V
- (D4) if $\dim(V) = \dim(W)$ and $V \subseteq W$, then $V = W$.

Proof.

- (D1) V is spanned by a finitely many vectors, by definition. Hence, we can always find k vectors that span V :

$$\text{Span}(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k) = V \quad (41.2.13)$$

If $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ are linearly independent, then we have found a basis.

Otherwise, one of the vectors can be expressed as a linear combination of the others, and can be dropped from the span. Repeat this process until the remaining set of vectors is linearly independent.

- (D2) Suppose $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ do not span V (since otherwise we would already have a basis). Then, $\exists \mathbf{v}_{k+1} \neq \text{Span}(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k)$, which can be added to our set of linearly independent vectors. Continue until V has been generated.
- (D3) If $\dim(V) = n$, then any basis of V must necessarily contain n vectors. Suppose a linearly independent set of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ does not form a basis. Then, $\exists \mathbf{v}_{n+1} \neq \text{Span}(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n)$ which can be added to the linearly independent set. However, by Proposition 34.9, any linearly independent set of m vectors must satisfy $m \leq n$, where n is the dimension of V . This would imply $n + 1 \leq n$, a contradiction.
- (D4) If $\dim V = \dim W = n$, then there exists a basis $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\} \subseteq V$. We can deduce using (D3) that $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\} \subseteq V$ must be a basis of W too, and thus:

$$\text{Span}(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n) = V = W \quad (41.2.14)$$

as desired. ■

41.3 Operations on subspaces

In this section we will more closely inspect the properties of vector subspaces, and the operations we can apply on them, namely sums, direct sums and direct products, as well as intersections and unions.

We begin by providing an alternative, often faster way to prove that some subsets are subspaces.

Proposition 34.12 (Subspace criterion)

For a subset $W \subseteq V$ to be a vector subspace of V over \mathbb{K} , we need:

$$c\mathbf{w}_1 + \mathbf{w}_2 \in W, \forall \mathbf{w}_1, \mathbf{w}_2 \in W, \forall c \in \mathbb{K} \quad (41.3.1)$$

Proof. We wish to prove that (S1)-(S3) are equivalent to 34.3.1. Firstly, note that:

$$(\mathbf{w}'_1 + \mathbf{w}'_2 \in W) \wedge (c\mathbf{w}'_1 \in W, \forall \mathbf{w}'_1, \mathbf{w}'_2 \in W, \forall c \in \mathbb{K}) \quad (41.3.2)$$

$$\implies c\mathbf{w}_1 + \mathbf{w}_2 \in W, \forall \mathbf{w}_1, \mathbf{w}_2 \in W, \forall c \in \mathbb{K} \quad (41.3.3)$$

if we take $\mathbf{w}'_1 = c\mathbf{w}_1$. Similarly:

$$c\mathbf{w}_1 + \mathbf{w}_2 \in W, \forall \mathbf{w}_1, \mathbf{w}_2 \in W, \forall c \in \mathbb{K} \quad (41.3.4)$$

$$\implies (\mathbf{w}'_1 + \mathbf{w}'_2 \in W) \wedge (c\mathbf{w}'_1 \in W, \forall \mathbf{w}'_1, \mathbf{w}'_2 \in W, \forall c \in \mathbb{K}) \quad (41.3.5)$$

if we take $c = 1$ and $\mathbf{w}_2 = \mathbf{w}_1$ to prove the left and right statements of 34.3.5 respectively.

Also:

$$c\mathbf{w}_1 + \mathbf{w}_2 \in W, \forall \mathbf{w}_1, \mathbf{w}_2 \in W, \forall c \in \mathbb{K} \implies \mathbf{0} \in W \quad (41.3.6)$$

if we take $c = -1$ and $\mathbf{w}_1 = \mathbf{w}_2$ ■

Theorem 34.13 (Spanning subspace)

Let $S \subseteq V$, then $\text{Span}(W)$ is a vector subspace of V .

Proof. Let $S = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k\} \subseteq V$. Then, $\mathbf{0} = \sum_{i=1}^k 0 \cdot \mathbf{u}_i \implies \mathbf{0} \in \text{Span}(W)$.

Now let $\mathbf{v}_1 = \sum_{i=1}^k \alpha_i \mathbf{u}_i \in \text{Span}(W)$ and $\mathbf{v}_2 = \sum_{i=1}^k \beta_i \mathbf{u}_i \in \text{Span}(W)$ with $\alpha_i, \beta_i \in \mathbb{K}$, then:

$$c\mathbf{v}_1 + \mathbf{v}_2 = c \sum_{i=1}^k \alpha_i \mathbf{u}_i + \sum_{i=1}^k \beta_i \mathbf{u}_i = \sum_{i=1}^k (c\alpha_i + \beta_i) \mathbf{u}_i = \sum_{i=1}^k \gamma_i \mathbf{u}_i \in \text{Span}(W) \quad (41.3.7)$$

where $\gamma_i = c\alpha_i + \beta_i \in \mathbb{K}$ due to the closure of fields. Hence, by Proposition 34.12, $\text{Span}(W)$ is a subspace of V . ■

Proposition 34.14 (Dimension of subspace)

The dimension of a vector subspace of V is always less than or equal to the dimension of V .

Proof. Let V be a vector space of dimension $\dim(V) = n$, and let $S \subseteq V$ be a subspace of V of dimension $\dim(W) = m$. Let $\mathcal{B}_V = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ be a basis of V , and $\mathcal{B}_W = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m\}$ be a basis of W , then it follows that \mathcal{B}_W is linearly independent. Hence, using Proposition 34.9, $\dim(W) = m < n = \dim(V)$. \blacksquare

Proposition 34.15 (Union and intersection of subspaces)

Let V be a finite dimensional vector space, with subspaces $\{W_1, W_2, \dots, W_n\}$. Then $\forall 1 \leq k \leq n$:

$$W = \bigcap_{i=1}^k W_i \text{ is a subspace of } V \quad \forall 1 \leq k \leq n \quad (41.3.8)$$

and for any two subspaces W_1, W_2 :

$$W_1 \cup W_2 \text{ is a subspace of } V \text{ iff } W_1 \subseteq W_2 \text{ or } W_2 \subseteq W_1 \quad (41.3.9)$$

Proof. We begin by proving that $\bigcap_{i=1}^k W_i$ is a subspace of V , $\forall 1 \leq k \leq n$. $\mathbf{0} \in W$, since $\mathbf{0} \in W_i$ for all i by the subgroup axioms.

Moreover, if $\mathbf{u}, \mathbf{v} \in W$, then $\mathbf{u}, \mathbf{v} \in W_i$ for all i . By proposition 34.12 then:

$$\alpha\mathbf{u} + \mathbf{v} \in W_i, \forall i \implies \alpha\mathbf{u} + \mathbf{v} \in W \quad (41.3.10)$$

as required. Hence, the subgroup criteria are met, and $W = \bigcap_{i=1}^k W_i$ is a subspace of V .

Next we prove that $W_1 \cup W_2$ is a subspace of V iff $W_1 \subseteq W_2$ or $W_2 \subseteq W_1$.

(\implies) We proceed by contradiction. Suppose $W_1 \cup W_2$ is a subspace of V and suppose $W_1 \not\subseteq V$ and $W_2 \not\subseteq V$. Then $\exists \mathbf{w}_1 \in W_1 \setminus W_2$ and $\exists \mathbf{w}_2 \in W_2 \setminus W_1$. Therefore, by the closure axiom of groups:

$$\mathbf{w}_1 + \mathbf{w}_2 \in V \quad (41.3.11)$$

Now suppose that $\mathbf{w}_1 + \mathbf{w}_2 \in W_1$. Then:

$$(-\mathbf{w}_1) + (\mathbf{w}_1 + \mathbf{w}_2) = \mathbf{w}_2 \in W_1 \quad (41.3.12)$$

which is a contradiction. So $\mathbf{w}_1 + \mathbf{w}_2 \notin W_1$. Similarly, suppose that $\mathbf{w}_1 + \mathbf{w}_2 \in W_2$. Then this would imply that:

$$(-\mathbf{w}_2) + (\mathbf{w}_1 + \mathbf{w}_2) = \mathbf{w}_1 \in W_2 \quad (41.3.13)$$

which is a contradiction. Thus, we conclude that $\mathbf{w}_1 + \mathbf{w}_2 \notin W_1 \cup W_2$. However, this violates the subspace criteria in Proposition 34.3. Hence, we must require $W_1 \subseteq W_2$ or $W_2 \subseteq W_1$.

(\Leftarrow) Suppose that $W_1 \subseteq W_2$ or $W_2 \subseteq W_1$. Then $W_1 \cup W_2 = W_2$ or $W_1 \cup W_2 = W_1$ respectively, and since both W_1, W_2 are both subspaces of W then it follows that $W_1 \cup W_2$ is in either cases a subspace of W .

■

Definition 34.16 (Cosets, quotient spaces, sums of spaces)

Let V be a vector space and let W be a vector subspace of V . Then, a coset of V is:

$$\mathbf{v} + W \equiv \{\mathbf{v} + \mathbf{w} : \forall \mathbf{v} \in V\} \quad (41.3.14)$$

where $\mathbf{w} \in W$. The set of all cosets of V in W is called the quotient space of W modulo V :

$$V/W \equiv \{\mathbf{v} + W : \forall \mathbf{v} \in V\} \quad (41.3.15)$$

Finally, the sum of two vector subspaces U, W of V is defined as:

$$U + W \equiv \{\mathbf{u} + \mathbf{w} : \forall \mathbf{u} \in U, \mathbf{w} \in W\} \quad (41.3.16)$$

Theorem 34.17 (Dimension of sum of spaces)

Let U, W be subspaces of a finite dimensional space V . Then:

$$\dim(U + W) = \dim(U) + \dim(W) - \dim(U \cap W). \quad (41.3.17)$$

Proof. We firstly prove that $\dim(U + W) = \dim(U) + \dim(W) - \dim(U \cap W)$.

Let $B_{\cap} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r\}$ be a basis for $U \cap W$ so that $\dim(U \cap W) = r$. From Proposition 34.15, we must have that $U \cup W$ is a subspace of both U and W , and consequently (D2) of Proposition 34.11 implies that B_{\cap} can be extended to form a basis of U and W .

Hence suppose that $B_U = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r, \mathbf{u}_1, \dots, \mathbf{u}_m\}$ and $B_W = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r, \mathbf{w}_1, \dots, \mathbf{w}_n\}$ are bases for U and W respectively.

Now let $\mathbf{v}' \in U + W$, then it may be expressed as:

$$\mathbf{v}' = \mathbf{u}' + \mathbf{w}' \quad (41.3.18)$$

$$= \underbrace{\left(\sum_{i=1}^r \alpha_i \mathbf{v}_i + \sum_{i=1}^m \alpha'_i \mathbf{u}_i \right)}_{\mathbf{u}'} + \underbrace{\left(\sum_{i=1}^r \beta_i \mathbf{v}_i + \sum_{i=1}^n \beta'_i \mathbf{w}_i \right)}_{\mathbf{w}'} \quad (41.3.19)$$

for some $\mathbf{u}' \in U, \mathbf{w}' \in W$. We can rearrange the above equation:

$$\mathbf{v}' = \left(\sum_{i=1}^r \alpha_i \mathbf{v}_i + \sum_{i=1}^m \alpha'_i \mathbf{u}_i \right) + \left(\sum_{i=1}^r \beta_i \mathbf{v}_i + \sum_{i=1}^n \beta'_i \mathbf{w}_i \right) \quad (41.3.20)$$

$$= \sum_{i=1}^r \gamma_i \mathbf{v}_i + \sum_{i=1}^m \alpha'_i \mathbf{u}_i + \sum_{i=1}^n \beta'_i \mathbf{w}_i \quad (41.3.21)$$

where $\gamma_i = \alpha_i + \beta_i$. Therefore:

$$\text{Span}(\mathbf{v}_1, \dots, \mathbf{v}_r, \mathbf{u}_1, \dots, \mathbf{u}_m, \mathbf{w}_1, \dots, \mathbf{w}_n) = U \quad (41.3.22)$$

Also, note that $\{\mathbf{v}_1, \dots, \mathbf{v}_r, \mathbf{u}_1, \dots, \mathbf{u}_m, \mathbf{w}_1, \dots, \mathbf{w}_n\}$ is linearly independent. Indeed, suppose:

$$\sum_{i=1}^r \alpha_i \mathbf{v}_i + \sum_{i=1}^m \beta_i \mathbf{u}_i + \sum_{i=1}^n \gamma_i \mathbf{w}_i = \mathbf{0} \implies \sum_{i=1}^r \alpha_i \mathbf{v}_i + \sum_{i=1}^m \beta_i \mathbf{u}_i = - \sum_{i=1}^n \gamma_i \mathbf{w}_i \quad (41.3.23)$$

for some $\alpha_i, \beta_i, \gamma_i \in \mathbb{K}$. The above belongs to U , since $\{\mathbf{v}_1, \dots, \mathbf{v}_r, \mathbf{u}_1, \dots, \mathbf{u}_m\}$ is a basis of U . Similarly, it also belongs to W . Therefore:

$$\sum_{i=1}^r \alpha_i \mathbf{v}_i + \sum_{i=1}^m \beta_i \mathbf{u}_i = - \sum_{i=1}^n \gamma_i \mathbf{w}_i \in U \cap W \quad (41.3.24)$$

and can therefore be written as:

$$\sum_{i=1}^r \alpha_i \mathbf{v}_i + \sum_{i=1}^m \beta_i \mathbf{u}_i = - \sum_{i=1}^n \gamma_i \mathbf{w}_i = \sum_{i=1}^r c_i \mathbf{v}_i \quad (41.3.25)$$

$$\implies \sum_{i=1}^m \beta_i \mathbf{u}_i = \sum_{i=1}^r d_i \mathbf{v}_i \quad \text{and} \quad - \sum_{i=1}^n \gamma_i \mathbf{w}_i = \sum_{i=1}^r c_i \mathbf{v}_i \quad (41.3.26)$$

where $d_i = c_i - \alpha_i$. Recall that \mathbf{u}_i and \mathbf{v}_i must be linearly independent, since they form a basis of U . Thus we obtain $\beta_i = 0$ and $d_i = 0 \implies c_i = \alpha_i$.

Similarly, we require that \mathbf{w}_i and \mathbf{v}_i be linearly independent since they form a basis for W . Consequently $\gamma_i = c_i = 0 \implies \alpha_i = 0$. Linear dependence is thus satisfied.

So we may claim that $\{\mathbf{v}_1, \dots, \mathbf{v}_r, \mathbf{u}_1, \dots, \mathbf{u}_m, \mathbf{w}_1, \dots, \mathbf{w}_n\}$ is a basis of $U + W$. It follows that:

$$\dim(U + W) = (r + m) + (r + n) - r = \dim(U) + \dim(W) - \dim(U \cap W) \quad (41.3.27)$$

as was desired. ■

Theorem 34.18 (Dimension of quotient spaces)

Let U, W be subspaces of a finite dimensional space V . Then:

$$\dim(V/W) = \dim(V) - \dim(W) \quad (41.3.28)$$

Proof. Let $\mathcal{B}_W = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n\}$ and $\mathcal{B}_V = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n, \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m\}$ be bases for W and V respectively. Let $\mathbf{v} + W \in V/W$, where $\mathbf{v} \in V$, then it may be expressed as:

$$\mathbf{v} + W = \sum_{i=1}^n \alpha_i \mathbf{w}_i + \sum_{i=1}^m \beta_i \mathbf{v}_i + V \quad (41.3.29)$$

$$= \sum_{i=1}^n \alpha_i (\mathbf{w}_i + W) + \sum_{i=1}^m \beta_i (\mathbf{v}_i + W) \quad (41.3.30)$$

$$= \sum_{i=1}^m \beta_i (\mathbf{v}_i + W) \quad (41.3.31)$$

$$\in \text{Span}(\mathbf{v}_1 + W, \mathbf{v}_2 + W, \dots, \mathbf{v}_m + W) \quad (41.3.32)$$

where $\sum_{i=1}^n \alpha_i (\mathbf{w}_i + W)$ disappears since it is equal to W , and can be reabsorbed into the second sum.

Also, we note that $\{\mathbf{v}_1 + W, \mathbf{v}_2 + W, \dots, \mathbf{v}_m + W\}$ is linearly independent. Indeed:

$$\sum_{i=1}^m \alpha_i (\mathbf{v}_i + W) = \mathbf{0} + W \equiv W \quad (41.3.33)$$

implies:

$$\sum_{i=1}^m \alpha_i \mathbf{v}_i \in W \implies \sum_{i=1}^m \alpha_i \mathbf{v}_i = \sum_{i=1}^n \beta_i \mathbf{w}_i \quad (41.3.34)$$

But $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n, \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m\}$ is linearly independent since it forms a basis for V . Consequently, $\alpha_i = 0$, thus proving linear independence of $\{\mathbf{v}_1 + W, \mathbf{v}_2 + W, \dots, \mathbf{v}_m + W\}$.

So we may claim that $\{\mathbf{v}_1 + W, \mathbf{v}_2 + W, \dots, \mathbf{v}_m + W\}$ is a basis for V/W . It follows immediately that:

$$\dim(V/W) = (m+n) - n = \dim(V) - \dim(W) \quad (41.3.35)$$

as required. ■

Definition 34.19 (Direct sums)

Let U, W be subspaces of a finite dimensional space V . Then, we say that V is the (internal) direct sum of U and W if:

$$(DS1) \quad U + W = V$$

$$(DS2) \quad U \cap W = \{\mathbf{0}\}$$

We then say that U and W are complementary spaces, and denote the direct sum as:

$$U \oplus W = V \quad (41.3.36)$$

Instead, given two arbitrary vector spaces V_1, V_2 then their (external) direct sum is defined as:

$$V_1 \oplus V_2 \equiv \{(\mathbf{v}_1, \mathbf{v}_2) : \mathbf{v}_1 \in V_1, \mathbf{v}_2 \in V_2\} \quad (41.3.37)$$

An immediate consequence of this definition is that:

$$\dim(U \oplus W) = \dim(U) + \dim(W) \quad (41.3.38)$$

or more generally that:

$$\dim\left(\bigoplus_{i=1}^n W_i\right) = \sum_{i=1}^n \dim(W)_i \quad (41.3.39)$$

Euclidean geometry in \mathbb{R}^3

42

Matrix algebra

43

Linear transformations

44.1 What is a map?

We begin by restating some common results on maps that you should be familiar with.

Definition 46.1 (Map, Domain, Image and Kernel)

A map between two sets X and Y assigns to each $x \in X$ some $y = f(x) \in Y$ referred to as the **image of x under f** :

$$f : X \rightarrow Y \quad (44.1.1)$$

$$x \mapsto f(x) \quad (44.1.2)$$

Here X is called the **domain** of f , denoted $\text{dom}(f)$. Instead, the set:

$$\text{Im}(f) = \{f(x) : x \in X\} \subseteq Y \quad (44.1.3)$$

is called the **image** of f . If f is a homomorphism, then the set:

$$\text{Ker}(f) = \{x \in X : f(x) = 0\} \quad (44.1.4)$$

is called the **kernel** of f .

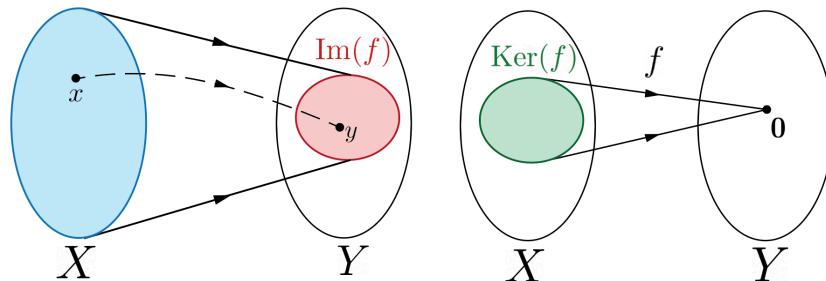


Figure 44.1. Map from X to Y .

Definition 46.2 (Injectivity, Surjectivity, Bijectivity) Recall that a function $f : X \rightarrow Y$ is said to be:

- (i) Injective: if every element of Y is the image of at most one element of X i.e.
 $f(x) = f(x') \implies x = x', \forall x, x' \in X$.
- (ii) Surjective: if every element of Y is the image of at least one element of X i.e.
 $\text{Im}(f) = Y$.
- (iii) Bijective: if it is both surjective and injective.

Recall that another way to state surjectivity is that if $\forall y \in Y, \exists x \in X s.t. f(x) = y$. This latter definition is equivalent to $y \in Y \implies y \in \text{Im}(f)$ so that $Y \subseteq \text{Im}(f)$. But $\text{Im}(f) \subseteq Y$ so $Y = \text{Im}(f)$.

As we saw in Group theory, it is possible to compose different elements of a dihedral group. Similarly, one can also compose maps.

Definition 46.3 (Map composition)

Given two maps $f : X \rightarrow Y$ and $g : Y \rightarrow Z$ then their composite map is defined as:

$$g \circ f : X \rightarrow Z \quad (44.1.5)$$

$$x \mapsto g(f(x)) \quad (44.1.6)$$

shown in the form of a commutative diagram below:

$$\begin{array}{ccc} A & \xrightarrow{f} & B \\ & \searrow^{g \circ f} & \downarrow g \\ & & C \end{array} \quad (44.1.7)$$

It is very important that $\text{Im}(f) \subseteq \text{dom } g$ since otherwise it would not be possible to evaluate $g(f(x))$. We can interpret the composite of two maps as another map which "jumps over" and bypasses Y as shown below:

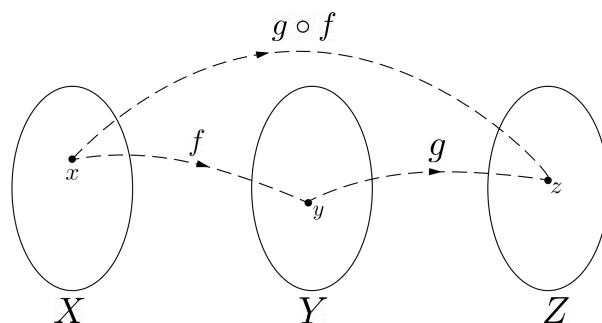


Figure 44.2. Composite map $g \circ f$

Suppose we wish to find two maps such that when composed together, they map back every element to itself. Such two maps would then be inverses of each other, so that when composed they give the identity transformation. We define them more rigorously below.

Definition 46.4 (Inverse and identity maps)

The identity map $\text{id}_X : X \rightarrow X$ maps all elements of X to themselves.

Given a map $f : X \rightarrow Y$, then $g : Y \rightarrow X$ is its inverse is:

$$(g \circ f) = \text{id}_X \quad \text{and} \quad (f \circ g) = \text{id}_Y \quad (44.1.8)$$

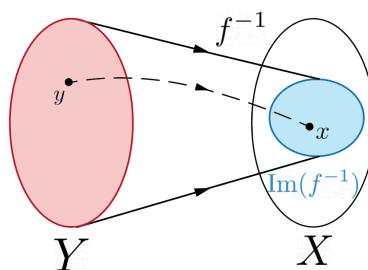


Figure 44.3. Inverse map f^{-1}

It is not always a given that a map f has an inverse. We can use the following theorem to determine which map f .

Theorem 46.4 (Bijectivity and invertibility)

The map $f : X \rightarrow Y$ has an inverse iff f is bijective, and this inverse is unique.

Proof.

(\implies) Suppose f has an inverse $g : Y \rightarrow X$. Then:

$$g \circ f = \text{id}_X \quad \text{and} \quad f \circ g = \text{id}_Y \quad (44.1.9)$$

If $f(x) = f(x')$ then $g(f(x)) = \text{id}_X(x) = x$ and $g(f(x')) = \text{id}_X(x') = x'$ so that $x = x'$, thus implying injectivity.

Let $y \in Y$. So $g(y) = x$ for some x , hence $f(g(y)) = \text{id}_Y(y) = y = f(x)$. Hence for any y there exists some x so that $y = f(x)$, giving surjectivity.

(\impliedby) Suppose f is bijective, so $\forall y \in Y$, there exists x s.t. $f(x) = y$. If we define $g(y) = x$ then for all $y \in Y$:

$$(f \circ g)(y) = f(x) = y \quad \text{and} \quad (g \circ f)(x) = g(y) = x \quad (44.1.10)$$

as required, g is its inverse.

Suppose f is invertible with two inverses g, h . Then:

$$(f \circ g)(x) = (f \circ h)(x) \forall x \in X \quad (44.1.11)$$

then composing with g $g(x) = h(x) \implies g = h$ since this equality holds for any x . ■

One can see this more intuitively. Indeed, if f is surjective, then there may be some elements in Y that are not mapped. Hence we cannot find an x so that $f^{-1}(y) = x$, which is clearly a problem since all y must get mapped by f^{-1} . If f is injective, then there may be several elements x mapping to the same y , so that $f^{-1}(y)$ is no longer well-defined. If however it is bijective, then every element of X gets mapped exactly once to some y , and all y are a map of some x , so invertibility is easily satisfied.

Proposition 46.6 (Important inverses)

Suppose f, g are bijective maps. Then f^{-1}, g^{-1} and $f \circ g$ are bijective, with inverses:

$$(f \circ g)^{-1} = g^{-1} \circ f^{-1} \text{ and } (f^{-1})^{-1} = f \quad (44.1.12)$$

Proof. The first follows from:

$$(f \circ g)^{-1} \circ (f \circ g) = \text{id} \quad (44.1.13)$$

and:

$$(g^{-1} \circ f^{-1}) \circ (f \circ g) = g^{-1} \circ \text{id} \circ g = \text{id} \quad (44.1.14)$$

so $(f \circ g)^{-1}$ and $g^{-1} \circ f^{-1}$ are inverses of $f \circ g$. But inverses are unique, hence the two must be equal.

Similarly:

$$(f^{-1})^{-1} \circ f^{-1} = \text{id} \text{ and } f \circ f^{-1} = \text{id} \quad (44.1.15)$$

hence by the same logic as before $f = (f^{-1})^{-1}$ as desired. ■

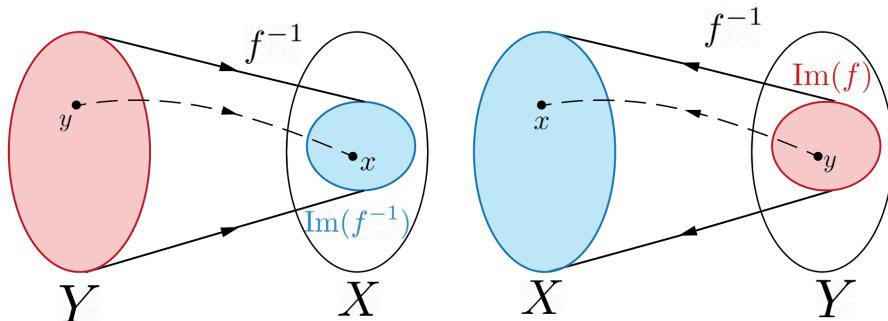


Figure 44.4. Left: diagram of f^{-1} treating it following definition 46.1 (so all elements of Y get mapped). Right: diagram of f^{-1} treating it as the map that undoes f

44.2 What is a linear map?

Definition 46.7 (*Linear map*)

A **linear map** is a map $f : V \rightarrow W$ between two vector spaces V and W over a field \mathbb{F} such that it preserves addition and scalar multiplication:

$$(L1) \quad f(\mathbf{v}_1 + \mathbf{v}_2) = f(\mathbf{v}_1) + f(\mathbf{v}_2), \quad \forall \mathbf{v}_1, \mathbf{v}_2 \in V$$

$$(L2) \quad f(\alpha \mathbf{v}_1) = \alpha f(\mathbf{v}_1), \quad \forall \mathbf{v}_1 \in V, \forall \alpha \in \mathbb{F}$$

The set of all linear maps from V to W is denoted $\text{Hom}(V, W)$

Proposition 46.8 (*Properties of linear maps*)

For any linear maps $f : V \rightarrow W$:

- (i) $f(\mathbf{0}) = \mathbf{0}$, that is, f fixes the zero vector.
- (ii) $\text{Ker}(f)$ is a subspace of V and $\text{Im}(f)$ is a subspace of W .
- (iii) f is surjective $\iff \text{Im}(f) = W \iff \dim \text{Im}(f) = \dim W$.
- (iv) f is injective $\iff \text{Ker}(f) = \{\mathbf{0}\} \iff \dim \text{Ker}(f) = 0$.
- (v) αf is linear for $\alpha \in \mathbb{K}$.
- (vi) if g is linear then $f + g$ is linear.
- (vii) if g is linear then $f \circ g$ is linear.

Proof.

- (i) We have already shown that $\mathbf{0} \in \text{Ker}(f)$ in point (i). $f(\mathbf{0}) = f(0 \cdot \mathbf{0}) = 0f(\mathbf{0}) = \mathbf{0}$
- (ii) Let $\mathbf{v}_1, \mathbf{v}_2 \in \text{Ker}(f)$. Then $f(\alpha \mathbf{v}_1 + \mathbf{v}_2) = \alpha f(\mathbf{v}_1) + f(\mathbf{v}_2) = \mathbf{0} \implies \alpha \mathbf{v}_1 + \mathbf{v}_2 \in \text{Ker}(f)$. Proposition 34.12 then ensures that $\text{Ker}(f)$ is a subspace of V .

We have already shown that $\mathbf{0} \in \text{Im}(f)$ in point (i). Let $\mathbf{w}_1, \mathbf{w}_2 \in \text{Im}(f)$, so that $\exists \mathbf{v}_1, \mathbf{v}_2 \in V$ such that $\mathbf{w}_i = f(\mathbf{v}_i)$. Then $\alpha \mathbf{w}_1 + \mathbf{w}_2 = \alpha f(\mathbf{v}_1) + f(\mathbf{v}_2) = f(\alpha \mathbf{v}_1 + \mathbf{v}_2) \in \text{Im}(f)$. Proposition 34.12 then ensures that $\text{Im}(f)$ is a subspace of W .

- (iii) If f is surjective, then $\forall \mathbf{w} \in W, \exists \mathbf{v} \in V$ such that $\mathbf{w} = f(\mathbf{v}) \in \text{Im}(f)$. So $W \subseteq \text{Im}(f)$, and $\text{Im}(f) \subseteq W$ implying $\text{Im}(f) \subseteq W$.

If instead $\text{Im}(f) = W$, then $W \subseteq \text{Im}(f)$, and $\text{Im}(f) \subseteq W$. Hence $\forall \mathbf{w} \in W, \exists \mathbf{v} \in V$ such that $\mathbf{w} = f(\mathbf{v}) \in \text{Im}(f)$, implying that f is surjective. It follows then that the two spaces have the same dimension.

- (iv) Suppose f is injective, and let $\mathbf{v}_1, \mathbf{0} \in \text{Ker}(f)$. Then $f(\mathbf{v}_1 - \mathbf{0}) = \mathbf{0}$ since $\text{Ker}(f)$ is a subspace. Consequently $f(\mathbf{v}_1) = f(\mathbf{0})$ and so $\mathbf{v}_1 = \mathbf{0}$. So $\text{Ker}(f) = \{\mathbf{0}\}$.

Suppose $\text{Ker}(f) = \{\mathbf{0}\}$, and let $\mathbf{v}_1, \mathbf{v}_2 \in \text{Ker}(f)$. Then, $f(\mathbf{v}_1) = f(\mathbf{v}_2) \implies f(\mathbf{v}_1 - \mathbf{v}_2) = \mathbf{0}$. Hence $\mathbf{v}_1 - \mathbf{v}_2 \in \text{Ker}(f)$ and so $\mathbf{v}_1 = \mathbf{v}_2$.

It follows that $\dim \text{Ker}(f) = \dim \{\mathbf{0}\} = 0$.



Definition 46.9 (Rank and nullity)

The dimension of the image of a linear map f is called its **rank**:

$$\text{rk}(f) = \dim \text{Im}(f) \quad (44.2.1)$$

and the dimension of the kernel of a linear map f is called its **nullity**:

$$\text{null}(f) = \dim \text{Ker}(f) \quad (44.2.2)$$

Theorem 46.10 (Rank-nullity theorem)

For any linear map $f : V \rightarrow W$:

$$\text{null}(f) + \text{rk}(f) = \dim(V) \quad (44.2.3)$$

Proof. Let $n = \dim V$ and $k = \text{null}(f)$, and let $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ be a basis for $\text{Ker}(f)$, which we can complete to $\{\mathbf{v}_1, \dots, \mathbf{v}_k, \mathbf{v}_{k+1}, \dots, \mathbf{v}_n\}$ to form a basis for V . We wish to show that $f(\mathbf{v}_{k+1} \dots \mathbf{v}_n)$ is a basis of $\text{Im}(f)$.

Let us firstly prove that $\text{Im}(f) = \text{Span}(f(\mathbf{v}_{k+1} \dots \mathbf{v}_n))$. Indeed consider $\mathbf{w} \in \text{Im}(f)$. Then there exists some \mathbf{v} such that:

$$\mathbf{w} = f(\mathbf{v}) = f\left(\sum_{i=1}^n \alpha_i \mathbf{v}_i\right) = \sum_{i=1}^n \alpha_i f(\mathbf{v}_i) \quad (44.2.4)$$

However, for $i \leq k$ we have that $f(\mathbf{v}_i) = \mathbf{0}$ and so $\text{Im}(f) = \text{Span}(f(\mathbf{v}_{k+1} \dots \mathbf{v}_n))$.

Now let us prove that $\{\mathbf{v}_{k+1} \dots \mathbf{v}_n\}$ are linearly independent. Consider:

$$\sum_{i=k+1}^n \alpha_i f(\mathbf{v}_i) = \mathbf{0} \implies f\left(\sum_{i=k+1}^n \alpha_i \mathbf{v}_i\right) = \mathbf{0} \quad (44.2.5)$$

and consequently $\sum_{i=k+1}^n \alpha_i \mathbf{v}_i \in \text{Ker}(f)$. Using the fact that $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ is a basis of $\text{Ker}(f)$ we find that:

$$\sum_{i=k+1}^n \alpha_i \mathbf{v}_i = \sum_{i=1}^k \beta_i \mathbf{v}_i \implies \sum_{i=1}^n \gamma_i \mathbf{v}_i = \mathbf{0} \quad (44.2.6)$$

where $\gamma_i = \alpha_i$ for $k < i \leq n$, and $\gamma_i = \beta_i$ for $1 \leq i \leq k$. However, we know that $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ forms a basis, and must therefore be linearly independent. Hence $\gamma_i = 0$ and so $\alpha_i = 0$ as well.

Since $\{\mathbf{v}_{k+1} \dots \mathbf{v}_n\}$ both generate $\text{Im}(f)$ and are also linearly independent, they must form a basis of $\text{Im}(f)$. So $\text{rk}(f) = n - k$ and consequently:

$$\dim V = n = k + (n - k) = \text{null}(f) + \text{rk}(f) \quad (44.2.7)$$

■

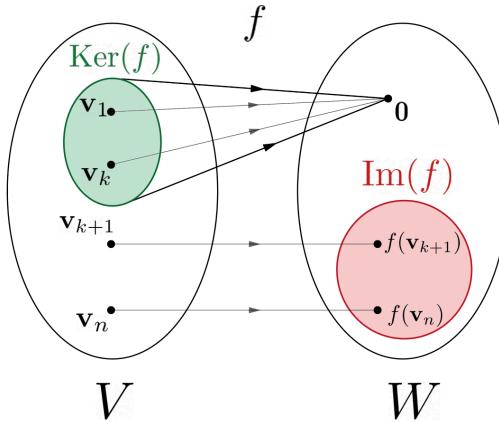


Figure 44.5. Visual depiction of the rank-nullity theorem

Throughout this proof we have also demonstrated how to construct a basis for $\text{Im}(f)$, which is to simply take the image of the basis of the domain.

It is then easy to see that for f to be invertible/bijective, then we need $\text{Ker}(f)$ to only contain $\mathbf{0}$, and no other vectors. In other words, we need $\text{null}(f) = 0$ and hence $\text{rk}(f) = \dim W$.

Let's justify this more rigorously.

Proposition 46.11 (Consequence of rank-nullity)

Let $f : V \rightarrow W$ be a linear transformation with $n = \dim V$. Then:

- (i) f is bijective $\implies \dim V = \dim W$.
- (ii) f is bijective $\iff \text{null}(f) = 0 \iff \text{rk}(f) = n$.
- (iii) If it exists, $f^1 : W \rightarrow V$ is linear.

Proof.

- (i) if f is bijective, then we showed in Proposition 46.8 that $\text{null}(f) = 0$ and $\text{rk}(f) = \dim W$. Using the rank nullity theorem:

$$\dim V = \text{rk}(f) = \dim W \quad (44.2.8)$$

as required.

- (ii) Suppose $\dim W = \dim V = n$ (which follows from bijectivity), then from the rank-nullity theorem:

$$n = \dim V = \text{rk}(f) + \text{null}(f) = \dim W \quad (44.2.9)$$

and since f is bijective then:

$$\text{null}(f) = 0 \implies \text{rk}(f) = n \quad (44.2.10)$$

(iii) Set $\mathbf{w}_1 = f(\mathbf{v}_1)$ and $\mathbf{w}_2 = f(\mathbf{v}_2)$. Then:

$$f^{-1}(\alpha\mathbf{w}_1 + \mathbf{w}_2) = f^{-1}(\alpha f(\mathbf{v}_1) + f(\mathbf{v}_2)) \quad (44.2.11)$$

$$= f^{-1}(f(\alpha\mathbf{v}_1 + \mathbf{v}_2)) \quad (44.2.12)$$

$$= \alpha\mathbf{v}_1 + \mathbf{v}_2 \quad (44.2.13)$$

$$= \alpha f^{-1}(\mathbf{w}_1) + f^{-1}(\mathbf{w}_2) \quad (44.2.14)$$

■

Proposition 46.12 (Linear map given basis of domain)

Let V and W be vector spaces, with $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ as a basis of V . For $\mathbf{w}_1, \dots, \mathbf{w}_n \in W$ there exists one linear map $f : V \rightarrow W$ such that $f(\mathbf{v}_i) = \mathbf{w}_i$.

Proof. One such linear map is:

$$f(\mathbf{v}) = \sum \alpha_i \mathbf{w}_i \quad (44.2.15)$$

where α_i are the coefficients of \mathbf{v}_i :

$$\mathbf{v} = \sum \alpha_i \mathbf{v}_i \quad (44.2.16)$$

It is linear since:

$$f(\beta\mathbf{v} + \mathbf{v}') = f(\sum \beta \alpha_i \mathbf{v}_i + \sum \alpha'_i \mathbf{v}_i) \quad (44.2.17)$$

$$= f\left(\sum (\beta \alpha_i + \alpha'_i) \mathbf{v}_i\right) \quad (44.2.18)$$

$$= \sum (\beta \alpha_i + \alpha'_i) \mathbf{w}_i \quad (44.2.19)$$

$$= \beta \sum \alpha_i \mathbf{w}_i + \sum \alpha'_i \mathbf{w}_i \quad (44.2.20)$$

Obviously, it maps $\mathbf{v}_i \mapsto \mathbf{w}_i$. Finally, it is unique, since if g was also a linear map with such properties then for all $\mathbf{v} \in V$:

$$g(\mathbf{v}) = \sum \alpha_i \mathbf{w}_i = f(\mathbf{v}) \implies f = g \quad (44.2.21)$$

as required. ■

44.3 Isomorphisms

Definition 46.13 (*Isomorphism*)

Two vector spaces V and W are said to be **isomorphic** whenever we can find an invertible linear map $f : V \rightarrow W$, called an **isomorphism** of V onto W .

It follows immediately from (i) of Proposition 46.11 that if two spaces are isomorphic, then they must have the same dimension. It turns out that the converse is also true.

Theorem 46.14 (*Equivalent statement of isomorphicity*)

Two finite dimensional vector spaces V and W are isomorphic *iff* $\dim V = \dim W$.

Proof. We have already proven \implies .

Now suppose that $\dim V = \dim W$ and let $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ and $\{\mathbf{w}_1, \dots, \mathbf{w}_n\}$ be bases for V and W respectively. Then we know from Proposition 46.12 that there exists a linear map f mapping the basis of V to the basis of W . Let us show that it is invertible, that is, bijective. Firstly, it is injective, since:

$$f(\mathbf{v}) = f(\mathbf{v}') \implies \sum \alpha_i \mathbf{v}_i = \sum \beta_i \mathbf{v}_i \implies \alpha_i = \beta_i \quad (44.3.1)$$

and so $\mathbf{v} = \mathbf{v}'$. It is also surjective since $\text{Im}(f) = \{f(\mathbf{v}_i) : i \in \mathbb{N}\} = \{\mathbf{w}_i : i \in \mathbb{N}\}$. Hence f is invertible, provided $\dim V = \dim W$. ■

44.4 Linear maps and matrices

Definition 46.15 (*Coordinate map*)

For a vector space V over \mathbb{K} , endowed with a basis $\beta = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$, then we define the **coordinate map** $\varphi : V \mapsto \mathbb{K}^n$ by:

$$\varphi_\beta(\mathbf{v}) = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix} = [\mathbf{v}]_\beta, \quad \forall \mathbf{v} \in V \quad (44.4.1)$$

where $\mathbf{v} = \sum_{i=1}^n \alpha_i \mathbf{v}_i$. Here, $[\mathbf{v}]_\beta$ is known as the **coordinate vector of \mathbf{v} relative to β** .

The inverse of φ_β is ϕ_β which given a basis β associates to a list of scalars $\alpha_1, \alpha_2, \dots, \alpha_n$ a vector \mathbf{v} .

Note that φ and ϕ are clearly linear. Moreover, $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ forms a basis of V then $\dim V = n = \dim \mathbb{K}^n$ implying that f is bijective and invertible. Consequently φ is an isomorphism, giving us the next insightful result:

Proposition 46.16 (Isomorphisms of V onto \mathbb{K}^n)

Every n dimensional vector space V is isomorphic to \mathbb{F}^n through a coordinate map ϕ .

Therefore, given a vector space V and a basis α then the maps ϕ_α and φ_α :

$$\mathbb{K}^n \xrightarrow{\phi_\alpha} V \xrightarrow{\varphi_\alpha} \mathbb{K}^n$$

Definition 46.17 (Matrix representation)

Let us now consider a basis $\beta = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ for V and a basis $\gamma = \{\mathbf{w}_1, \dots, \mathbf{w}_m\}$ for W . Let f be a linear transformation. Then, there exist unique scalars $a_{ij} \in \mathbb{K}$ for $1 \leq i \leq m$ such that for each $1 \leq j \leq n$:

$$f(\mathbf{v}_j) = \sum_{i=1}^m a_{ij} \mathbf{w}_i, \quad (44.4.2)$$

The scalars a_{ij} form a matrix A of size $m \times n$, called the **matrix representation** of T in the bases β and γ , denoted as $A = [f]_{\beta}^{\gamma}$.

We can draw a **commutative diagram** to demonstrate how matrix representations work:

$$\begin{array}{ccc} V & \xrightarrow{f} & W \\ \phi_\beta \uparrow & & \uparrow \phi_\gamma \\ \mathbb{K}^n & \xrightarrow{[f]_{\beta}^{\gamma}} & \mathbb{K}^m \end{array}$$

In this diagram, we designate the fields over which we define V and W in the bottom row. These contain the coordinate vectors of any $\mathbf{v} \in V, \mathbf{w} \in W$. To map from \mathbb{K}^n to V given a basis β , we need ϕ_β as defined in Definition 46.15. Similarly to map from \mathbb{K}^m to W we need ϕ_γ . Finally, if $f : V \rightarrow W$ then we can map from \mathbb{K}^n to \mathbb{K}^m (map from one coordinate vector to another) by multiplying by the matrix representation.

Proposition 46.18 (Columns of matrix representation)

Given a matrix representation A of f in the bases β and γ then:

$$A = ([f(\mathbf{v}_1)]_{\gamma} \ [f(\mathbf{v}_2)]_{\gamma} \ \dots \ [f(\mathbf{v}_n)]_{\gamma}) \quad (44.4.3)$$

so that $f(\mathbf{v}) = A\mathbf{v}$ for all $\mathbf{v} \in V$ (or alternatively $[f(\mathbf{v})]_{\gamma} = A[\mathbf{v}]_{\beta}$).

Proof. Note that for $\mathbf{v} \in V$:

$$f(\mathbf{v}) = f\left(\sum_{j=1}^n v_j \mathbf{v}_j\right) = \sum_{j=1}^n v_j f(\mathbf{v}_j) \quad (44.4.4)$$

so that if we let $[f(\mathbf{v}_j)]_\gamma = (a_{1j} \ a_{2j} \ \dots \ a_{mj})^T$ then:

$$[f(\mathbf{v})]_\gamma = \begin{pmatrix} a_{11}v_1 + a_{12}v_2 + \dots + a_{1n}v_n \\ a_{21}v_1 + a_{22}v_2 + \dots + a_{2n}v_n \\ \vdots \\ a_{m1}v_1 + a_{m2}v_2 + \dots + a_{mn}v_n \end{pmatrix} \quad (44.4.5)$$

We can express this more compactly as:

$$[f(\mathbf{v})]_\gamma = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \quad (44.4.6)$$

$$= ([f(\mathbf{v}_1)]_\gamma \ [f(\mathbf{v}_2)]_\gamma \ \dots \ [f(\mathbf{v}_n)]_\gamma) [\mathbf{v}]_\beta \quad (44.4.7)$$

$$\implies [f(\mathbf{v})]_\gamma = \mathbf{A}[\mathbf{v}]_\beta \quad (44.4.8)$$

More abstractly, one may write:

$$f(\mathbf{v}) = f\left(\sum_{j=1}^n v_j \mathbf{v}_j\right) = \sum_{j=1}^n v_j \sum_{i=1}^m a_{ij} \mathbf{w}_i = \sum_{i=1}^m \left(\sum_{j=1}^n v_j a_{ij} \right) \mathbf{w}_i \quad (44.4.9)$$

so that we end up with:

$$(f(\mathbf{v}))_i = \sum_{j=1}^n v_j a_{ij} = (\mathbf{Av})_i \quad (44.4.10)$$

This however is the expression of matrix multiplication \mathbf{Av} , as desired. ■

Theorem 46.19 (*Uniqueness of matrix representation*)

The matrix representation of a linear representation with respect to the bases β and γ is unique.

Proof. Suppose we had two different matrix representations \mathbf{A} and \mathbf{B} as shown below:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \text{ and } \mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{pmatrix} \quad (44.4.11)$$

Then the image of any basis vector $\mathbf{v}_j \in \beta$ is from Proposition 46.18:

$$[f(\mathbf{v}_j)]_\gamma = (a_{1j} \ a_{2j} \ \dots \ a_{mj})^T = (b_{1j} \ b_{2j} \ \dots \ b_{mj})^T \quad (44.4.12)$$

so j th column of \mathbf{A} and \mathbf{B} coincide. Since this is true for any $1 \leq j \leq n$ we have that $\mathbf{A} = \mathbf{B}$. ■

Proposition 46.20 (Properties of matrix representations)

Let V and W be finite dimensional vector spaces with bases β and γ respectively, and let f, g be linear maps of V onto W . Then:

- (i) $[f + g]_\beta^\gamma = [f]_\beta^\gamma + [g]_\beta^\gamma$
- (ii) $[\alpha f]_\beta^\gamma = \alpha[f]_\beta^\gamma$

Proof. Let $\beta = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ and $\gamma = \{\mathbf{w}_1, \dots, \mathbf{w}_m\}$ then there exists unique a_{ij} and b_{ij} such that:

$$f(\mathbf{v}_j) = \sum_{i=1}^m a_{ij} \mathbf{w}_i, \quad \text{and} \quad g(\mathbf{v}_j) = \sum_{i=1}^m b_{ij} \mathbf{w}_i \quad (44.4.13)$$

Then:

$$(\alpha f + g)(\mathbf{v}_j) = \sum_{i=1}^n (\alpha a_{ij} + b_{ij}) \mathbf{w}_i \quad (44.4.14)$$

and so:

$$([f + g]_\beta^\gamma)_{ij} = \alpha a_{ij} + b_{ij} = \alpha[f]_\beta^\gamma + [g]_\beta^\gamma \quad (44.4.15)$$

■

Theorem 46.21 (Matrix representation of composition)

Let V, W, U be finite dimensional vector spaces with bases α, β and γ respectively. Then, if $g : V \rightarrow W$ and $f : W \rightarrow U$ are linear maps:

$$[f \circ g]_\alpha^\gamma = [f]_\beta^\gamma [g]_\alpha^\beta \quad (44.4.16)$$

Proof. Let $C = [f \circ g]_\alpha^\gamma$, $A = [f]_\beta^\gamma$, $B = [g]_\alpha^\beta$. Also, let $\alpha = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$, $\beta = \{\mathbf{w}_1, \dots, \mathbf{w}_m\}$, $\gamma = \{\mathbf{u}_1, \dots, \mathbf{u}_l\}$. We can draw the following commutative diagram:

$$\begin{array}{ccccc} V & \xrightarrow{g} & W & \xrightarrow{f} & U \\ \phi_\alpha \uparrow & & \phi_\beta \uparrow & & \phi_\gamma \uparrow \\ \mathbb{K}^n & \xrightarrow{[g]_\alpha^\beta} & \mathbb{K}^m & \xrightarrow{[f]_\beta^\gamma} & \mathbb{K}^l \end{array}$$

Then for $\mathbf{v}_i \in \alpha$:

$$(f \circ g)(\mathbf{v}_i) = f\left(\sum_{k=1}^m B_{ki} \mathbf{w}_k\right) = \sum_{k=1}^m B_{ki} f(\mathbf{w}_k) \quad (44.4.17)$$

$$= \sum_{k=1}^m B_{ki} \left(\sum_{j=1}^l A_{jk} \mathbf{u}_k \right) \quad (44.4.18)$$

$$= \sum_{k=1}^m \sum_{j=1}^l A_{jk} B_{ki} \mathbf{u}_k = \sum_{k=1}^l C_{ji} \mathbf{u}_k \quad (44.4.19)$$

where:

$$C_{ji} = \sum_{j=1}^l A_{jk} B_{ki} = \mathbf{A}_j \cdot \mathbf{B}^i \quad (44.4.20)$$

However, this is exactly the definition of matrix multiplication we encountered in the previous chapter, hence:

$$[f \circ g]_\alpha^\gamma = \mathbf{C} = \mathbf{AB} = [f]_\beta^\gamma [g]_\alpha^\beta \quad (44.4.21)$$

■

Proposition 46.22 (*Invertibility of linear maps*)

Let $f : V \rightarrow W$ be a linear map, where α and β are bases of V and W respectively.

Then f is invertible iff $[f]_\beta^\gamma$ is invertible, then $[f^{-1}]_\beta^\gamma = ([f]_\beta^\gamma)^{-1}$.

Proof. Suppose that f has an inverse, so that $\dim V = \dim W = n$. Then $[f]_\beta^\gamma$ is a square matrix of size n , and:

$$\mathbb{I}_n = [\text{id}_V]_\beta = [f^{-1} \circ f]_\beta = [f^{-1}]_\gamma^\beta [f]_\beta^\gamma \quad (44.4.22)$$

Similarly $[f]_\beta^\gamma [f^{-1}]_\gamma^\beta$, so $[f]_\beta^\gamma$ is invertible, and $[f^{-1}]_\beta^\gamma = ([f]_\beta^\gamma)^{-1}$.

Now suppose $\mathbf{A} = [f]_\beta^\gamma$ is invertible, then there exists \mathbf{B} so that $\mathbf{AB} = \mathbf{BA} = \mathbb{I}_n$. Then there exists a map $g \in \text{Hom}(W, V)$ such that:

$$g(\mathbf{w}_j) = \sum_{i=1}^n B_{ij} \mathbf{v}_i \quad (44.4.23)$$

where $\gamma = \{\mathbf{w}_1, \dots, \mathbf{w}_n\}$ and $\beta = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ are bases of W and V respectively. Then:

$$[g \circ f]_\beta = [g]_\gamma^\beta [f]_\beta^\gamma = \mathbf{BA} = \mathbb{I}_n = [\text{id}_V]_\beta \quad (44.4.24)$$

so we conclude that $g \circ f = \text{id}_V$ (and similarly $f \circ g = \text{id}_V$) by the uniqueness of matrix representations.

■

Theorem 46.23 (All linear maps have a representation)

The vector spaces $\text{Mat}_{n,m}(\mathbb{K})$ and $\text{Hom}(V, W)$ are isomorphic.

Proof. Let $\beta = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ and $\gamma = \{\mathbf{w}_1, \dots, \mathbf{w}_m\}$ be bases for V and W respectively. Then:

$$\Phi : \text{Hom}(V, W) \rightarrow \text{Mat}_{n,m}(\mathbb{K}) \quad (44.4.25)$$

$$f \mapsto [f]_{\beta}^{\gamma} \quad (44.4.26)$$

is an isomorphism. Indeed, it is linear, as shown in proposition 46.20. Furthermore, it is injective, since matrix representations are unique as was shown in Theorem 46.19. Finally, it is surjective, since given any matrix \mathbf{A} then we can always find a linear map such that:

$$f(\mathbf{v}_i) = \sum_{i=1}^m A_{ij} \mathbf{w}_j \quad (44.4.27)$$

as warranted by Proposition 46.12. But this implies that $\Phi(f) = [f]_{\beta}^{\gamma} = \mathbf{A}$ as required. ■

An immediate consequence is that given two vector spaces V, W of dimensions n and m then $\text{Hom}(V, W)$ has dimension $n \cdot m$.

This is a fundamental result since it proves that given any matrix, we can associate it to some linear map. Similarly, to every linear map we can associate some matrix representations. **Generally, a map is linear iff it has a matrix representation.**

Since every matrix defines a linear map, we can define its nullity and rank.

Definition 46.24 (Matrix rank)

Given a matrix $\mathbf{A} : \mathbb{F}^n \rightarrow \mathbb{F}^m$, its associated linear map is:

$$\mathbf{Av} = \sum_{i=1}^m v_i f(\mathbf{v}_i) = \sum_{i=1}^m v_i \mathbf{A}^i \quad (44.4.28)$$

then the image of \mathbf{A} is $\text{Im}(\mathbf{A}) = \text{Span}(\mathbf{A}^1, \dots, \mathbf{A}^m)$. Its column rank, $\text{rk}(\mathbf{A})$, is the number of linearly independent columns of \mathbf{A} . Its row rank is the number of linearly independent rows of \mathbf{A} , so $\text{rk}(\mathbf{A}^T)$.

Proposition 46.25 (Column rank and row rank)

Column rank and row rank are the same.

Proof. Suppose that \mathbf{A}_1 can be written as a linear combination:

$$\mathbf{A}_1 = \sum_{j=1}^n \alpha_j \mathbf{A}_j \quad (44.4.29)$$

and let $\alpha = (\alpha_2 \ \alpha_3 \ \dots \ \alpha_n)^T$. Let:

$$\mathbf{A}^i = \begin{pmatrix} a_i \\ \mathbf{b}_i \end{pmatrix}, \quad \mathbf{b}_i = \begin{pmatrix} b_{2i} \\ b_{3i} \\ \vdots \\ b_{ni} \end{pmatrix} \implies \mathbf{A} = \begin{pmatrix} a_1 & a_2 & \dots & a_m \\ b_{21} & b_{22} & \dots & b_{2m} \\ b_{31} & b_{32} & \dots & b_{3m} \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & \dots & b_{nm} \end{pmatrix} \quad (44.4.30)$$

Consequently

$$a_i = (\mathbf{A}^i)_1 = (\mathbf{A}_1)_i = \sum_{j=2}^n \alpha_j A_{ji} = \sum_{j=2}^n \alpha_j (\mathbf{A}^i)_j \quad (44.4.31)$$

$$= \alpha_2 b_{2i} + \alpha_3 b_{3i} + \dots + \alpha_n b_{ni} \quad (44.4.32)$$

$$= \boldsymbol{\alpha} \cdot \mathbf{b}_i \quad (44.4.33)$$

so that

$$\mathbf{A}^i = \begin{pmatrix} \boldsymbol{\alpha} \cdot \mathbf{b}_i \\ \mathbf{b}_i \end{pmatrix} \quad (44.4.34)$$

Therefore, if one row is a linear combination of the others, then we can drop it, leaving the row rank unchanged obviously. However, the column rank also remains unchanged. Indeed dropping the first row of \mathbf{A} we get:

$$\mathbf{A} = \begin{pmatrix} \boldsymbol{\alpha} \cdot \mathbf{b}_1 & \boldsymbol{\alpha} \cdot \mathbf{b}_2 & \dots & \boldsymbol{\alpha} \cdot \mathbf{b}_m \\ b_{21} & b_{22} & \dots & b_{2m} \\ b_{31} & b_{32} & \dots & b_{3m} \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & \dots & b_{nm} \end{pmatrix} \longrightarrow \mathbf{A}' = \begin{pmatrix} b_{21} & b_{22} & \dots & b_{2m} \\ b_{31} & b_{32} & \dots & b_{3m} \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & \dots & b_{nm} \end{pmatrix} \quad (44.4.35)$$

the column rank remains the same. Indeed, the first element of any column already contains a linear combination of all elements below it. Consequently, the column spans of \mathbf{A} and \mathbf{A}' are generated by the same number of elements. In other words, the column rank is unchanged by removing a linearly dependent row.

More rigorously, we need to prove that removing the row containing $\boldsymbol{\alpha} \cdot \mathbf{b}_i$ will not alter the linear dependence of the columns. Suppose that $\mathbf{A}^1, \mathbf{A}^2, \dots, \mathbf{A}^l$ is the maximal linearly independent set of columns of \mathbf{A} . Then note that

$$c_1 \begin{pmatrix} \boldsymbol{\alpha} \cdot \mathbf{b}_1 \\ \mathbf{b}_1 \end{pmatrix} + c_2 \begin{pmatrix} \boldsymbol{\alpha} \cdot \mathbf{b}_2 \\ \mathbf{b}_2 \end{pmatrix} + \dots + c_l \begin{pmatrix} \boldsymbol{\alpha} \cdot \mathbf{b}_l \\ \mathbf{b}_l \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha} \cdot (c_1 \mathbf{b}_1 + \dots + c_l \mathbf{b}_l) \\ c_1 \mathbf{b}_1 + \dots + c_l \mathbf{b}_l \end{pmatrix} \quad (44.4.36)$$

So if $\mathbf{A}^1, \mathbf{A}^2, \dots, \mathbf{A}^l$ are linearly independent, then $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_l$ are also linearly independent. Moreover, adding any other \mathbf{b}_i will result in linear dependence. Indeed, if this were not

the case, then $c'_1 \mathbf{b}_1 + \dots + c'_{l+1} \mathbf{b}_{l+1} \implies c'_i = 0$ and so:

$$\begin{pmatrix} \alpha \cdot (c'_1 \mathbf{b}_1 + \dots + c'_{l+1} \mathbf{b}_{l+1}) \\ c'_1 \mathbf{b}_1 + \dots + c'_{l+1} \mathbf{b}_{l+1} \end{pmatrix} = c'_1 \begin{pmatrix} \alpha \cdot \mathbf{b}_1 \\ \mathbf{b}_1 \end{pmatrix} + \dots + c'_{l+1} \begin{pmatrix} \alpha \cdot \mathbf{b}_{l+1} \\ \mathbf{b}_{l+1} \end{pmatrix} = \mathbf{0} \implies c'_i = 0 \quad (44.4.37)$$

so the columns $\mathbf{A}^1, \mathbf{A}^2, \dots, \mathbf{A}^{l+1}$ would be linearly independent, a contradiction.

A similar reasoning can be used to show that the row rank is unchanged by removing a linearly dependent column.

So, if we continue this process, removing linearly dependent rows/columns, eventually we will end up with a final matrix \mathbf{A} , whose row and column ranks will not have been altered, and whose rows and columns will be linearly independent. This final matrix must be forcibly square. If it were $n \times m$, assume WLOG $n < m$ then the m vectors in \mathbb{F}^n must be linearly dependent, a contradiction. Hence, the number of linearly independent rows and columns are the same, as desired. ■

44.5 Change of basis and equivalence

Definition 46.26 (Standard basis)

The standard basis of a coordinate vector space \mathbb{K}^n is:

$$\mathbf{e}_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \leftarrow i\text{ith component} \quad (44.5.1)$$

and its coordinate map is denoted φ_{id} .

Suppose we are given the coordinate vector of some vector $\mathbf{v} \in V$ in the **standard coordinates**:

$$\mathbf{v} = \sum_{i=1}^n v_i \mathbf{e}_i \quad (44.5.2)$$

How do we find its coordinates in some other basis $\beta = \mathbf{v}_1, \dots, \mathbf{v}_n$?

Consider the commutative diagram below:

$$\begin{array}{ccc} V & \xleftarrow{\text{id}_V} & V \\ \phi_{\text{id}} \uparrow & & \uparrow \phi_{\beta} \\ \mathbb{K}_n & \xleftarrow{\mathbf{P}_{\beta}} & \mathbb{K}_n \end{array}$$

then:

$$P_\beta = [\text{id}_V]_\beta^{\text{id}} = ([\mathbf{v}_1]_{\text{id}} \ [\mathbf{v}_2]_{\text{id}} \ \dots \ [\mathbf{v}_n]_{\text{id}}) \quad (44.5.3)$$

whereas:

$$M_\beta = P_\beta^{-1} = [\text{id}_V]_{\text{id}}^\beta = ([\mathbf{e}_1]_\beta \ [\mathbf{e}_2]_\beta \ \dots \ [\mathbf{e}_n]_\beta) \quad (44.5.4)$$

Indeed, notice that:

$$M_\beta[\mathbf{v}]_{\text{id}} = ([\mathbf{e}_1]_\beta \ [\mathbf{e}_2]_\beta \ \dots \ [\mathbf{e}_n]_\beta)[\mathbf{v}]_{\text{id}} \quad (44.5.5)$$

$$= v_1[\mathbf{e}_1]_\beta + v_2[\mathbf{e}_2]_\beta + \dots + v_n[\mathbf{e}_n]_\beta \quad (44.5.6)$$

$$= [v_1\mathbf{e}_1 + v_2\mathbf{e}_2 + \dots + v_n\mathbf{e}_n]_\beta \quad (44.5.7)$$

$$= [\mathbf{v}]_\beta = [\text{id}_V(\mathbf{v})]_\beta \quad (44.5.8)$$

where v_1, v_2, \dots, v_n are the components of $[\mathbf{v}]_{\text{id}}$. Similarly:

$$P_\beta[\mathbf{v}]_\beta = ([\mathbf{v}_1]_{\text{id}} \ [\mathbf{v}_2]_{\text{id}} \ \dots \ [\mathbf{v}_n]_{\text{id}})[\mathbf{v}]_\beta \quad (44.5.9)$$

$$= v'_1[\mathbf{v}_1]_{\text{id}} + v'_2[\mathbf{v}_2]_{\text{id}} + \dots + v'_n[\mathbf{v}_n]_{\text{id}} \quad (44.5.10)$$

$$= [v'_1\mathbf{v}_1 + v'_2\mathbf{v}_2 + \dots + v'_n\mathbf{v}_n]_{\text{id}} \quad (44.5.11)$$

$$= [\mathbf{v}]_{\text{id}} = [\text{id}_V(\mathbf{v})]_{\text{id}} \quad (44.5.12)$$

where v'_1, v'_2, \dots, v'_n are the components of $[\mathbf{v}]_\beta$.

Definition 46.27 (*Transition matrix*)

If $\beta = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ is a basis of a vector space V then:

$$M_\beta = ([\mathbf{e}_1]_\beta \ [\mathbf{e}_2]_\beta \ \dots \ [\mathbf{e}_n]_\beta) \quad (44.5.13)$$

called the **transition matrix** maps $\varphi_{\text{id}}(\mathbf{v}) \mapsto \varphi_\beta(\mathbf{v})$ so that:

$$[\mathbf{v}]_\beta = M_\beta[\mathbf{v}]_{\text{id}} \quad (44.5.14)$$

We can view the transition matrix as the **matrix representation** of φ_β . Similarly P_β is the matrix representation of ϕ_β .

Let us generalize this result for any two bases:

Proposition 46.28 (*Change of coordinate matrix*)

Let $\beta = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ and γ be two bases of a vector space V . Then:

$$M_{\beta \rightarrow \gamma} = ([\mathbf{v}_1]_\gamma \ [\mathbf{v}_2]_\gamma \ \dots \ [\mathbf{v}_n]_\gamma) \quad (44.5.15)$$

is the **change of coordinate matrix** mapping $\varphi_\beta(\mathbf{v}) \mapsto \varphi_\gamma(\mathbf{v})$ so that:

$$[\mathbf{v}]_\gamma = M_{\beta \rightarrow \gamma}[\mathbf{v}]_\beta \quad (44.5.16)$$

Proof. Consider the following commutative diagram:

$$\begin{array}{ccccc}
 & & M_{\beta \rightarrow \gamma} & & \\
 & \swarrow P_\beta & & \searrow M_\gamma & \\
 \mathbb{K}_n & \xrightarrow{\quad} & \mathbb{K}_n & \xrightarrow{\quad} & \mathbb{K}_n \\
 \downarrow \phi_\beta & & \downarrow \phi_{\text{id}} & & \downarrow \phi_\gamma \\
 V & \xrightarrow{\text{id}_V} & V & \xrightarrow{\text{id}_V} & V
 \end{array}$$

so that it is clear that:

$$M_{\beta \rightarrow \gamma} = [\varphi_\gamma \circ \phi_\beta]_{\text{id}} = [\text{id}]_\beta^\gamma \quad (44.5.17)$$

Then

$$M_{\beta \rightarrow \gamma} = M_\gamma P_\beta = M_\gamma([\mathbf{v}_1]_{\text{id}} \ [\mathbf{v}_2]_{\text{id}} \ \dots \ [\mathbf{v}_n]_{\text{id}}) \quad (44.5.18)$$

$$= (\phi_\gamma([\mathbf{v}_1]_{\text{id}}) \ \phi_\gamma([\mathbf{v}_2]_{\text{id}}) \ \dots \ \phi_\gamma([\mathbf{v}_n]_{\text{id}})) \quad (44.5.19)$$

$$= ([\mathbf{v}_1]_\gamma \ [\mathbf{v}_2]_\gamma \ \dots \ [\mathbf{v}_n]_\gamma) \quad (44.5.20)$$

as desired. Alternatively, using proposition 46.18 with $f = \text{id}_V$ then:

$$M_{\beta \rightarrow \gamma} = [\text{id}]_\beta^\gamma = ([\mathbf{v}_1]_\gamma \ [\mathbf{v}_2]_\gamma \ \dots \ [\mathbf{v}_n]_\gamma) \quad (44.5.21)$$

as found previously. ■

Proposition 46.29 (*Invertibility of change of coordinate matrix*)

The inverse of $M_{\beta \rightarrow \gamma}$ is $M_{\beta \rightarrow \gamma}^{-1} = M_{\gamma \rightarrow \beta}$.

Proof. Let us firstly prove that $M_{\gamma \rightarrow \beta} M_{\beta \rightarrow \gamma} = \mathbb{I}_n$. Indeed, let $\beta = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ so that:

$$M_{\gamma \rightarrow \beta} M_{\beta \rightarrow \gamma} = [\text{id}_V]_\gamma^\beta [\text{id}_V]_\beta^\gamma = [\text{id}_V]_\beta^\beta = ([\mathbf{v}_1]_\beta \ [\mathbf{v}_2]_\beta \ \dots \ [\mathbf{v}_n]_\beta) = \mathbb{I}_n \quad (44.5.22)$$

where we used Theorem 46.21. Similarly, one can also find that $M_{\beta \rightarrow \gamma} M_{\gamma \rightarrow \beta}$. ■

Theorem 46.30 (*Change of basis of linear transformation*)

Let β, β' be two bases for V and let γ, γ' be two bases for W . If $f : V \rightarrow W$ is a linear map, then:

$$A' = M_{\gamma \rightarrow \gamma'} A M_{\beta' \rightarrow \beta} \quad (44.5.23)$$

Proof. Consider the commutative diagram below:

$$\begin{array}{ccccccc}
 V & \xrightarrow{\text{id}_V} & V & \xrightarrow{f} & W & \xleftarrow{\text{id}_W} & W \\
 \phi_{\beta'} \uparrow & & \phi_\beta \uparrow & & \phi_\gamma \uparrow & & \phi_{\gamma'} \uparrow \\
 \mathbb{K}^m & \xrightarrow{M_{\beta' \rightarrow \beta}} & \mathbb{K}^m & \xrightarrow{A} & \mathbb{K}^n & \xleftarrow{M_{\gamma' \rightarrow \gamma}} & \mathbb{K}^n \\
 & \searrow A' & & & \swarrow & & \\
 & & & & & &
 \end{array}$$

Then note that:

$$A = \varphi_\gamma \circ f \circ \phi_\beta \quad (44.5.24)$$

$$A' = \varphi_{\gamma'} \circ f \circ \phi_{\beta'} \quad (44.5.25)$$

hence:

$$A' = \varphi_{\gamma'} \circ f \circ \phi_{\beta'} \quad (44.5.26)$$

$$= \varphi_{\gamma'} \circ (\phi_\gamma \circ \varphi_\gamma) \circ f \circ (\phi_\beta \circ \varphi_\beta) \circ \phi_{\beta'} \quad (44.5.27)$$

$$= (\varphi_{\gamma'} \circ \phi_\gamma) \circ A \circ (\varphi_\beta \circ \phi_\beta) \quad (44.5.28)$$

$$= M_{\gamma' \rightarrow \gamma}^{-1} A M_{\beta' \rightarrow \beta} \quad (44.5.29)$$

but we have proven that $M_{\gamma' \rightarrow \gamma}^{-1} = M_{\gamma \rightarrow \gamma'}$ so that:

$$A' = M_{\gamma \rightarrow \gamma'} A M_{\beta' \rightarrow \beta} \quad (44.5.30)$$

as desired. ■

We can interpret this result with some intuition by considering the action of each matrix in 46.5.29 on $[\mathbf{v}]_{\beta'}$. Indeed, the matrix $M_{\beta' \rightarrow \beta}$ converts it to $[\mathbf{v}]_\beta$. Then A maps it to $A[\mathbf{v}]_\beta = [f(\mathbf{v})]_\gamma$ by Proposition 46.18. Finally $M_{\gamma \rightarrow \gamma'}$ maps it to $[f(\mathbf{v})]_{\gamma'}$.

Definition 46.31 (Equivalent matrices)

Let $A, B \in \text{Mat}_{n,m}(\mathbb{K})$ are **equivalent matrices** if there exists invertible matrices $P, Q \in \text{Mat}_m(\mathbb{K})$ such that:

$$B = Q^{-1} A P \quad (44.5.31)$$

We see immediately that if two matrices represent the same linear map with respect to different bases, then they are similar.

Proposition 46.32 (Similarity to special matrix)

Any matrix $A \in \text{Mat}_{m,n}(\mathbb{K})$ is equivalent to:

$$\begin{pmatrix} \mathbb{I}_r & 0 \\ 0 & 0 \end{pmatrix} \quad (44.5.32)$$

where $r = \text{rk}(A)$.

Proof. We begin by proving that any linear map $f : V \rightarrow W$ has some set of bases β and γ of V and W respectively such that:

$$[f]_{\beta}^{\gamma} = \begin{pmatrix} \mathbb{I}_r & 0 \\ 0 & 0 \end{pmatrix} \quad (44.5.33)$$

Set r so that $\text{null}(f) = n - r$, so that $\ker f$ has basis $\mathbf{v}_{r+1}, \dots, \mathbf{v}_n$ which we extend to $\mathbf{v}_1, \dots, \mathbf{v}_r, \mathbf{v}_{r+1}, \dots, \mathbf{v}_n$ a basis for V . We know that $f(\mathbf{v}_1), f(\mathbf{v}_2), \dots, f(\mathbf{v}_n)$ is a basis for $\text{Im}(f)$, and can be extended to a basis γ for W . Consequently:

$$[f]_{\beta}^{\gamma} = ([f(\mathbf{v}_1)]_{\gamma} [f(\mathbf{v}_2)]_{\gamma} \dots [f(\mathbf{v}_r)]_{\gamma} [f(\mathbf{v}_{r+1})]_{\gamma} \dots [f(\mathbf{v}_n)]_{\gamma}) \quad (44.5.34)$$

but $f(\mathbf{v}_i) = \mathbf{0}$ for $r < i \leq n$ so that:

$$[f]_{\beta}^{\gamma} = ([f(\mathbf{v}_1)]_{\gamma} [f(\mathbf{v}_2)]_{\gamma} \dots [f(\mathbf{v}_r)]_{\gamma} \mathbf{0} \dots \mathbf{0}) = \begin{pmatrix} \mathbb{I}_r & 0 \\ 0 & 0 \end{pmatrix} \quad (44.5.35)$$

Now, if we choose some matrix $A \in \text{Mat}_{m,n}(\mathbb{K})$, then by Theorem 46.23 it must be the matrix representation of some linear map f . Hence, it must be equivalent to:

$$\begin{pmatrix} \mathbb{I}_r & 0 \\ 0 & 0 \end{pmatrix} \quad (44.5.36)$$

which is also another representation of f as desired. ■

We can use this theorem to prove the equivalence of row and column rank. Indeed, if we let $A \in \text{Mat}_{m,n}(\mathbb{K})$ then there exist $Q, P \in \text{Mat}_m(\mathbb{K})$ such that:

$$Q^{-1}AP = \begin{pmatrix} \mathbb{I}_r & 0 \\ 0 & 0 \end{pmatrix} \quad (44.5.37)$$

implying that:

$$(Q^{-1}AP)^T = P^T A^T (Q^{-1})^T = \begin{pmatrix} \mathbb{I}_r & 0 \\ 0 & 0 \end{pmatrix} \quad (44.5.38)$$

so that A^T is also equivalent to $\begin{pmatrix} \mathbb{I}_r & 0 \\ 0 & 0 \end{pmatrix}$, and hence A^T and A both represent the same map and must therefore have the same rank. So column rank and row rank are the same.

Solving linear equations

45.1 Structure of solutions

Let's consider a linear map $f : V \rightarrow W$. Suppose we wish to find the solutions to

$$f(\mathbf{x}) = \mathbf{b}, \quad \mathbf{x} \in V, \mathbf{b} \in W \quad (45.1.1)$$

For $\mathbf{b} \neq 0$, this is known as the inhomogeneous linear equation, whose associated homogeneous equation is:

$$f(\mathbf{x}) = \mathbf{0} \quad (45.1.2)$$

Clearly, the solution of the latter is $\text{Ker}(f)$.

Proposition (Structure of solutions)

Suppose $\mathbf{x}_0 \in V$ is a solution to the inhomogeneous equation (45.1.1). Then the general solution is given by:

$$\mathbf{x} = \mathbf{x}_0 + \text{Ker}(f) \quad (45.1.3)$$

Proof. Suppose $\mathbf{x} \in V$ is a general solution to (45.1.1) so that $f(\mathbf{x}) = \mathbf{b}$. Since \mathbf{x}_0 is a solution, we have that $f(\mathbf{x}_0) = \mathbf{b}$, so we may write that $f(\mathbf{x} - \mathbf{x}_0) = \mathbf{0}$. This implies that $\mathbf{x} - \mathbf{x}_0 \in \text{Ker}(f)$ or alternatively that $\mathbf{x} = \mathbf{x}_0 + \text{Ker}(f)$. ■

Suppose we have a linear transformation f represented by a matrix $\mathbf{A} : \mathbb{K}^n \rightarrow \mathbb{K}^m$. Then, for $\mathbf{x} \in \mathbb{K}^n$ and $\mathbf{b} \in \mathbb{K}^m$ we consider the linear system of equations $\mathbf{Ax} = \mathbf{b}$:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2 \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n = b_m \end{cases} \quad (45.1.4)$$

We must now consider the following three scenarios:

- (i) if $\text{rk}(\mathbf{A}) = m$, then a solution exists for any choice of \mathbf{b} . Indeed, since $\text{Im}(\mathbf{A}) = \mathbb{K}^m$ (since $\text{Im}(\mathbf{A}) \subseteq \mathbb{K}^m$ and they have the same dimensionality), it follows that any vector

$\mathbf{b} \in \mathbb{K}^m$ may be expressed as an image of f . The number of free parameters is given by $\dim \text{Ker}(\mathbf{A}) = n - m$. Geometrically, if we let $m = n = 3$ and $\mathbb{K} = \mathbb{R}$, then we see that if $\text{rk}(\mathbf{A}) = 3$ then the linear map f maps the typical Euclidean space to itself.

- (ii) if $\text{rk}(\mathbf{A}) < m$ and $\mathbf{b} \in \text{Im}(\mathbf{A})$ then solution exists. Indeed, in this case a generic case of \mathbf{b} will no longer work, we must be careful and make sure that it belongs to the image of \mathbf{A} . The number of free parameters will be $\dim \text{Ker}(\mathbf{A}) = n - \text{rk}(\mathbf{A})$. Geometrically, this corresponds to the linear map f mapping the Euclidean space to a subspace of itself, such as a plane. We may add any vector that maps to the origin to a solution. For example in the case where $\text{Im}(\mathbf{A})$ is a plane the kernel will be a line, any vector on this line may be added, giving a free parameter.
- (iii) if $\text{rk}(\mathbf{A}) < m$ but $\mathbf{b} \notin \text{Im}(\mathbf{A})$ then solution doesn't exist. Indeed if the vector \mathbf{b} doesn't lie in the space spanned by \mathbf{A} then clearly a solution will not exist, since no vector \mathbf{x} will get mapped to \mathbf{b} .

Definition (Augmented matrix)

Consider the linear system of equations $\mathbf{Ax} = \mathbf{b}$ where $\mathbf{A} : \mathbb{K}^n \rightarrow \mathbb{K}^m$ is a matrix and $\mathbf{x} \in \mathbb{K}^n$ and $\mathbf{b} \in \mathbb{K}^m$. We define its augmented matrix to be:

$$(\mathbf{A}|\mathbf{b}) = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ \vdots & \vdots & & \vdots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} & b_m \end{array} \right) \quad (45.1.5)$$

Theorem (Rank of augmented matrix)

For a matrix $\mathbf{A} : \mathbb{K}^n \rightarrow \mathbb{K}^m$ and $\mathbf{b} \in \mathbb{K}^n$:

$$\mathbf{b} \in \text{Im}(\mathbf{A}) \iff \text{rk}(\mathbf{A}) = \text{rk}((\mathbf{A}|\mathbf{b})) \quad (45.1.6)$$

Proof. \implies Suppose that $\mathbf{b} \in \text{Im}(\mathbf{A})$. Then adding this vector to \mathbf{A} will not alter the space it spans, so that $\text{rk}(\mathbf{A}) = \text{rk}((\mathbf{A}|\mathbf{b}))$.

\impliedby Suppose that $\text{rk}(\mathbf{A}) = \text{rk}((\mathbf{A}|\mathbf{b}))$. Then, this means that:

$$\text{Span}(\mathbf{A}^1, \mathbf{A}^2, \dots, \mathbf{A}^n, \mathbf{b}) = \text{Span}(\mathbf{A}^1, \mathbf{A}^2, \dots, \mathbf{A}^n) \quad (45.1.7)$$

showing that $\mathbf{b} \in \text{Im}(\mathbf{A})$. ■

Therefore, if we can find the rank of the augmented matrix and show that it is equal to the rank of the coefficient matrix \mathbf{A} , then we have shown that a solution must indeed exist. We can find the rank of matrices using elementary matrix operations.

45.2 Elementary matrix operations

Definition (*Elementary row operations*)

The following are **elementary row operations**:

(R1) exchange two rows

(R2) scale a row by a non-zero scalar

(R3) add a non-zero multiple of a row to another row

These operations do not alter the rank of a matrix.

Because these operations do not alter the rank of a matrix, we may use them to transform a given matrix into a simpler one where it is easier to determine the span of its column/row vectors.

Definition ((Reduced) row echelon form)

A matrix is said to be in **row echelon form** if:

- (i) a leading entry in a non-zero row is strictly to the right of the leading entry in the row above
- (ii) zero rows are at the bottom

so it has general form:

$$\begin{pmatrix} \dots & a_{ij_1} & \dots & \dots & \dots & \dots & * \\ \vdots & & a_{2j_2} & & & & \vdots \\ \vdots & & & & & & \vdots \\ \vdots & & & & & a_{rj_r} & \dots \\ \mathbf{0} & \mathbf{0} & \dots & \dots & \dots & \dots & \mathbf{0} \end{pmatrix} \quad (45.2.1)$$

Instead, a matrix is said to be in **reduced row echelon form** if:

- (i) it is in row echelon form
- (ii) each leading entry is a 1
- (iii) each leading 1 is the only non-zero entry in its column

For example, the following matrix:

$$\begin{pmatrix} 0 & 1 & 0 & 2 & 0 & 7 \\ 0 & 0 & 1 & -3 & 0 & 2 \\ 0 & 0 & 0 & 0 & 1 & 7 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad (45.2.2)$$

is not in reduced row echelon form since the leading 1 in the fourth row is not the only non-zero element in its column. However it is in row echelon form.

Instead, the following matrix:

$$\begin{pmatrix} 1 & 0 & 0 & 3 \\ 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad (45.2.3)$$

is indeed in reduced row echelon form.

This gives us a strategy, known as Gauss-Jordan elimination to solve systems of linear equations.

Strategy (Gauss-Jordan elimination)

- (i) Apply row operations to $(A|b)$ until the matrix A within it is in row-echelon form.
- (ii) Let $r = \text{rk}(A)$. If $b_i \neq 0$ for some $i > r$, then this means that $\text{rk}(A) \neq \text{rk}(A|b)$, and hence $b \notin \text{Im}(A)$. The system therefore has no solutions.
- (iii) Otherwise, convert to reduced row echelon form, and solve the resulting system of equations.

Example. Consider:

$$\begin{cases} 3x_1 - 11x_2 - 3x_3 = 3 \\ 2x_1 - 6x_2 - 2x_3 = 1 \\ 5x_1 - 17x_2 - 6x_3 = 2 \\ 4x_1 - 8x_2 = 7 \end{cases} \quad (45.2.4)$$

We construct the augmented matrix:

$$(A|b) = \left(\begin{array}{ccc|c} 3 & -11 & -3 & 3 \\ 2 & -6 & -2 & 1 \\ 5 & -17 & -6 & 2 \\ 4 & -8 & 0 & 7 \end{array} \right) \quad (45.2.5)$$

which we reduce to row echelon form:

$$(A|b) = \left(\begin{array}{ccc|c} 3 & -11 & -3 & 3 \\ 2 & -6 & -2 & 1 \\ 5 & -17 & -6 & 2 \\ 4 & -8 & 0 & 7 \end{array} \right) \rightarrow \left(\begin{array}{ccc|c} 3 & -11 & -3 & 3 \\ 0 & \frac{4}{3} & 0 & -1 \\ 0 & \frac{4}{3} & -1 & -3 \\ 0 & \frac{20}{3} & 4 & 3 \end{array} \right) \quad (45.2.6)$$

$$\rightarrow \left(\begin{array}{ccc|c} 3 & -11 & -3 & 3 \\ 0 & \frac{4}{3} & 0 & -1 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 4 & 8 \end{array} \right) \rightarrow \left(\begin{array}{ccc|c} 3 & 0 & -3 & -\frac{21}{4} \\ 0 & \frac{4}{3} & 0 & -1 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & 0 \end{array} \right) \quad (45.2.7)$$

$$(45.2.8)$$

So we see that $\text{rk}(\mathbf{A}) = \text{rk}((\mathbf{A}|\mathbf{b})) = 3$ implying that $\mathbf{b} \in \text{Im}(\mathbf{A})$, and that a solution to the system exists.

We therefore convert the augmented matrix into reduced row echelon form:

$$(\mathbf{A}|\mathbf{b}) \rightarrow \left(\begin{array}{ccc|c} 3 & 0 & -3 & -\frac{21}{4} \\ 0 & \frac{4}{3} & 0 & -1 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & 0 \end{array} \right) \rightarrow \left(\begin{array}{ccc|c} 3 & 0 & 0 & \frac{3}{4} \\ 0 & \frac{4}{3} & 0 & -1 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & 0 \end{array} \right) \quad (45.2.9)$$

$$\rightarrow \left(\begin{array}{ccc|c} 1 & 0 & 0 & \frac{1}{4} \\ 0 & 1 & 0 & -\frac{3}{4} \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{array} \right) \quad (45.2.10)$$

which gives the solution:

$$x_1 = \frac{1}{4}, \quad x_2 = -\frac{3}{4}, \quad x_3 = 2 \quad (45.2.11)$$

◀

A useful tool to check whether or not an arithmetic mistake has been made while performing row reduction is to write the sum of the entries in each row as an extra column, so for example. For example

$$\left(\begin{array}{ccc|c|c} 3 & -11 & -3 & 3 & -8 \\ 0 & \frac{4}{3} & 0 & -1 & \frac{1}{3} \\ 0 & \frac{4}{3} & -1 & -3 & -\frac{8}{3} \\ 0 & \frac{20}{3} & 4 & 3 & \frac{41}{3} \end{array} \right) \quad (45.2.12)$$

When we perform a row operation, we perform it on this extra column as well. If the numbers in this final column still correspond to the sum of the elements in the corresponding row, then no mistakes have been made. In our example, we get:

$$\left(\begin{array}{ccc|c|c} 3 & -11 & -3 & 3 & -8 \\ 0 & \frac{4}{3} & 0 & -1 & \frac{1}{3} \\ 0 & 0 & -1 & -2 & -3 \\ 0 & 0 & 4 & 8 & 12 \end{array} \right) \quad (45.2.13)$$

so we do indeed find that the sum of all the rows are given in the transformed fifth column.

If say we had gotten, say, 13 in the last row, then an arithmetic mistake must have been made.

45.3 Inverting matrices

Definition (*Elementary matrix*)

The matrices obtained by performing row operations on $\mathbb{1}$ are known as elementary matrices.

Note that elementary matrices are important because they represent row operations. Suppose we have some row operation, which when acted on $\mathbb{1}_n$ gives the elementary matrix E . Then applying the same row operation on another $n \times n$ matrix A we will get $A' = EA$. For example, if E represents the exchange of rows i and j , then:

$$\mathbb{1} = (\mathbf{e}_1 \dots \mathbf{e}_j \dots \mathbf{e}_i \dots \mathbf{e}_n)^T \implies E = (\mathbf{e}_1 \dots \mathbf{e}_j \dots \mathbf{e}_i \dots \mathbf{e}_n)^T \quad (45.3.1)$$

so that:

$$E\mathbb{A} = (\mathbf{e}_1 \dots \mathbf{e}_j \dots \mathbf{e}_i \dots \mathbf{e}_n)^T (\mathbf{A}_1 \dots \mathbf{A}_i \dots \mathbf{A}_j \dots \mathbf{A}_n) \quad (45.3.2)$$

$$= \begin{pmatrix} \mathbf{e}_1 \cdot \mathbf{A}_1 & \mathbf{e}_1 \cdot \mathbf{A}_2 & \dots & \mathbf{e}_1 \cdot \mathbf{A}_n \\ \vdots & \vdots & & \vdots \\ \mathbf{e}_j \cdot \mathbf{A}_1 & \mathbf{e}_j \cdot \mathbf{A}_2 & \dots & \mathbf{e}_j \cdot \mathbf{A}_n \\ \vdots & \vdots & & \vdots \\ \mathbf{e}_i \cdot \mathbf{A}_1 & \mathbf{e}_i \cdot \mathbf{A}_2 & \dots & \mathbf{e}_i \cdot \mathbf{A}_n \\ \vdots & \vdots & & \vdots \\ \mathbf{e}_n \cdot \mathbf{A}_1 & \mathbf{e}_n \cdot \mathbf{A}_2 & \dots & \mathbf{e}_n \cdot \mathbf{A}_n \end{pmatrix} \quad (45.3.3)$$

which is indeed the version of A with the i th and j th rows exchanged.

$$A = \mathbb{1}A = (\mathbf{e}_1 \dots \mathbf{e}_j \dots \mathbf{e}_i \dots \mathbf{e}_n)^T (\mathbf{A}_1 \dots \mathbf{A}_i \dots \mathbf{A}_j \dots \mathbf{A}_n) \quad (45.3.4)$$

$$= \begin{pmatrix} \mathbf{e}_1 \cdot \mathbf{A}_1 & \mathbf{e}_1 \cdot \mathbf{A}_2 & \dots & \mathbf{e}_1 \cdot \mathbf{A}_n \\ \vdots & \vdots & & \vdots \\ \mathbf{e}_i \cdot \mathbf{A}_1 & \mathbf{e}_i \cdot \mathbf{A}_2 & \dots & \mathbf{e}_i \cdot \mathbf{A}_n \\ \vdots & \vdots & & \vdots \\ \mathbf{e}_j \cdot \mathbf{A}_1 & \mathbf{e}_j \cdot \mathbf{A}_2 & \dots & \mathbf{e}_j \cdot \mathbf{A}_n \\ \vdots & \vdots & & \vdots \\ \mathbf{e}_n \cdot \mathbf{A}_1 & \mathbf{e}_n \cdot \mathbf{A}_2 & \dots & \mathbf{e}_n \cdot \mathbf{A}_n \end{pmatrix} \quad (45.3.5)$$

Similarly, if E_2 represents multiplication of the i th row by a scalar λ then clearly:

$$E_2\mathbb{A} = (\mathbf{e}_1 \dots \lambda \mathbf{e}_i \dots \mathbf{e}_n)^T (\mathbf{A}_1 \dots \mathbf{A}_i \dots \mathbf{A}_n) \quad (45.3.6)$$

$$= \begin{pmatrix} \mathbf{e}_1 \cdot \mathbf{A}_1 & \mathbf{e}_1 \cdot \mathbf{A}_2 & \dots & \mathbf{e}_1 \cdot \mathbf{A}_n \\ \vdots & \vdots & & \vdots \\ \lambda \mathbf{e}_i \cdot \mathbf{A}_1 & \lambda \mathbf{e}_i \cdot \mathbf{A}_2 & \dots & \lambda \mathbf{e}_i \cdot \mathbf{A}_n \\ \vdots & \vdots & & \vdots \\ \mathbf{e}_n \cdot \mathbf{A}_1 & \mathbf{e}_n \cdot \mathbf{A}_2 & \dots & \mathbf{e}_n \cdot \mathbf{A}_n \end{pmatrix} \quad (45.3.7)$$

which is the version of A with the i th row multiplied by λ .

Finally, suppose that E_3 represents adding a λ -multiple of the j th row to the i th row. Then we find that:

$$E_1 A = (\mathbf{e}_1 \dots \mathbf{e}_i \dots \mathbf{e}_j + \lambda \mathbf{e}_i \dots \mathbf{e}_n)^T (A_1 \dots A_i \dots A_j \dots A_n) \quad (45.3.8)$$

$$= \begin{pmatrix} \mathbf{e}_1 \cdot A_1 & \mathbf{e}_1 \cdot A_2 & \dots & \mathbf{e}_1 \cdot A_n \\ \vdots & \vdots & & \vdots \\ \mathbf{e}_i \cdot A_1 & \mathbf{e}_i \cdot A_2 & \dots & \mathbf{e}_i \cdot A_n \\ \vdots & \vdots & & \vdots \\ \mathbf{e}_j \cdot A_1 + \lambda \mathbf{e}_i \cdot A_1 & \mathbf{e}_j \cdot A_2 + \lambda \mathbf{e}_i \cdot A_2 & \dots & \mathbf{e}_j \cdot A_n + \lambda \mathbf{e}_i \cdot A_n \\ \vdots & \vdots & & \vdots \\ \mathbf{e}_n \cdot A_1 & \mathbf{e}_n \cdot A_2 & \dots & \mathbf{e}_n \cdot A_n \end{pmatrix} \quad (45.3.9)$$

which is indeed the version of A with the λ -multiple of the i th row added to the j th row.

Note also that elementary row matrices are all invertible, because the row operations they represent are all invertible.

Theorem (Invertibility theorem)

- (a) An $n \times n$ square matrix A is invertible iff its reduced row echelon form is $\mathbb{1}$, so if $\text{rk}(A) = n$.
- (b) Any sequence of row operations that transform A to $\mathbb{1}$ also transform $\mathbb{1}$ to A^{-1} .

Proof. Let A be an $n \times n$ matrix whose reduced row echelon form is:

$$U = E_k E_{k-1} \dots E_1 A = BA \quad (45.3.10)$$

where E_k, E_{k-1}, \dots, E_1 are elementary matrices. Since they are all invertible, we have that B^{-1} exists.

\Rightarrow Suppose A is invertible. Then U is the product of invertible matrices, and is therefore invertible itself. Hence it cannot have any zero rows, since such matrices are not invertible. It follows from the conditions of the reduced row echelon form of matrices that the only possible choice of U is $\mathbb{1}$. Indeed, it is upper triangular (so in row echelon) with each leading entry as 1. Because it has n leading 1s and n rows, these leading entries must be on the diagonal. Finally, since there are no other entries on each column with a leading 1, and all columns have a leading 1, we get the identity matrix.

\Leftarrow Suppose $U = \mathbb{1}$, then:

$$BA = \mathbb{1} \implies A = B^{-1}\mathbb{1} \implies AB = \mathbb{1} \quad (45.3.11)$$

Note however that $\mathbb{1}$ and B^{-1} are invertible, so A will also be invertible, with $A^{-1} = B$.

Therefore, we find that:

$$A^{-1} = B = E_k E_{k-1} \dots E_1 \mathbb{1} \implies (A|\mathbb{1}) \rightarrow (\mathbb{1}|A^{-1}) \quad (45.3.12)$$

so we find A^{-1} by applying the same row operations that row reduce A to \mathbb{I} . ■

We can use the invertibility theorem to find the inverse of matrices.

Example. Let's find the inverse of the following matrix:

$$A = \begin{pmatrix} 1 & 4 & 1 \\ 1 & 6 & 3 \\ 2 & 3 & 0 \end{pmatrix} \quad (45.3.13)$$

We find that:

$$\left(\begin{array}{ccc|ccc} 1 & 4 & 1 & 1 & 0 & 0 \\ 1 & 6 & 3 & 0 & 1 & 0 \\ 2 & 3 & 0 & 0 & 0 & 1 \end{array} \right) \rightarrow \left(\begin{array}{ccc|ccc} 1 & 4 & 1 & 1 & 0 & 0 \\ 0 & 2 & 2 & -1 & 1 & 0 \\ 0 & -5 & -2 & -2 & 0 & 1 \end{array} \right) \rightarrow \left(\begin{array}{ccc|ccc} 1 & 4 & 1 & 1 & 0 & 0 \\ 0 & 2 & 2 & -1 & 1 & 0 \\ 0 & 0 & 3 & -\frac{9}{2} & \frac{5}{2} & 1 \end{array} \right) \quad (45.3.14)$$

$$\rightarrow \left(\begin{array}{ccc|ccc} 1 & 4 & 1 & 1 & 0 & 0 \\ 0 & 2 & 0 & 2 & -\frac{2}{3} & -\frac{2}{3} \\ 0 & 0 & 1 & -\frac{3}{2} & \frac{5}{6} & \frac{1}{3} \end{array} \right) \rightarrow \left(\begin{array}{ccc|ccc} 1 & 4 & 0 & \frac{5}{2} & -\frac{5}{6} & -\frac{1}{3} \\ 0 & 2 & 0 & 2 & -\frac{2}{3} & -\frac{2}{3} \\ 0 & 0 & 1 & -\frac{3}{2} & \frac{5}{6} & \frac{1}{3} \end{array} \right) \quad (45.3.15)$$

$$\rightarrow \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -\frac{3}{2} & \frac{1}{2} & 1 \\ 0 & 2 & 0 & 2 & -\frac{2}{3} & -\frac{2}{3} \\ 0 & 0 & 1 & -\frac{3}{2} & \frac{5}{6} & \frac{1}{3} \end{array} \right) \rightarrow \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -\frac{3}{2} & \frac{1}{2} & 1 \\ 0 & 1 & 0 & 1 & -\frac{1}{3} & -\frac{1}{3} \\ 0 & 0 & 1 & -\frac{3}{2} & \frac{5}{6} & \frac{1}{3} \end{array} \right) \quad (45.3.16)$$

so we see that:

$$A^{-1} = \frac{1}{6} \begin{pmatrix} -9 & 3 & 6 \\ 6 & -2 & -2 \\ -9 & 5 & 2 \end{pmatrix} \quad (45.3.17)$$

To check:

$$\frac{1}{6} \begin{pmatrix} -9 & 3 & 6 \\ 6 & -2 & -2 \\ -9 & 5 & 2 \end{pmatrix} \begin{pmatrix} 1 & 4 & 1 \\ 1 & 6 & 3 \\ 2 & 3 & 0 \end{pmatrix} = \frac{1}{6} \begin{pmatrix} 6 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 6 \end{pmatrix} = \mathbb{I} \quad (45.3.18)$$

as expected. ◀

The invertibility of matrices is especially important when solving linear systems of equations, since they can be used to find solutions whenever the associated system has only trivial solutions (that is, when the kernel of the matrix is null, and hence the rank is maximal).

Proposition (Linear systems and invertibility)

For an $n \times n$ matrix A , the following statements are equivalent:

- (a) A is invertible
- (b) The system $Ax = b$ has a unique solution for any b
- (c) The system $Ax = \mathbf{0}$ only has a trivial solution.

Proof.

- (a) \implies (b) Let A be invertible. Suppose $Ax = b$, then multiplying by A^{-1} then $A^{-1}Ax = A^{-1}b \implies x = A^{-1}b$. Instead, if $x = A^{-1}b$ then multiplying by A we get $Ax = b$.
- (b) \implies (c) Suppose $Ax = b$ has a unique solution for any b . Then $Ax = 0$ also has a unique solution, which can only be the trivial solution
- (c) \implies (a) Suppose that $Ax = 0$ only has a trivial solution. Then this means that when we reduce the augmented matrix:

$$\left(\begin{array}{ccc|c} a_{11} & a_{12} & a_{13} & 0 \\ a_{21} & a_{22} & a_{23} & 0 \\ a_{31} & a_{32} & a_{33} & 0 \end{array} \right) \quad (45.3.19)$$

then we must get that:

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array} \right) \quad (45.3.20)$$

since there can only be trivial solutions. So A has a reduced row echelon form of \mathbb{I} , proving by the Invertibility theorem that it is invertible.

Alternatively, we could note that if $\dim \text{Ker}(A) = 0$ then $\text{rk}(A) = n$. Therefore A is invertible. ■

Example. Let's consider the system:

$$\begin{cases} x + 4y + z = 4 \\ x + 6y + 3z = 6 \\ 2x + 3y = 9 \end{cases} \quad (45.3.21)$$

which may be written in matrix form as:

$$Ax = b, \quad A = \begin{pmatrix} 1 & 4 & 1 \\ 1 & 6 & 3 \\ 2 & 3 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 4 \\ 6 \\ 9 \end{pmatrix} \quad (45.3.22)$$

We have already found A^{-1} , so we find that:

$$x = A^{-1}b = \frac{1}{6} \begin{pmatrix} -9 & 3 & 6 \\ 6 & -2 & -2 \\ -9 & 5 & 2 \end{pmatrix} \begin{pmatrix} 4 \\ 6 \\ 9 \end{pmatrix} = \frac{1}{6} \begin{pmatrix} 36 \\ -6 \\ 12 \end{pmatrix} = \begin{pmatrix} 6 \\ -1 \\ 2 \end{pmatrix} \quad (45.3.23)$$

so the solution to the system of equations is $x = 6, y = -1, z = 2$. ◀

Proposition (Inverse matrix properties)

Let A, B be invertible matrices. Then:

- (a) $(A^T)^{-1} = (A^{-1})^T$
- (b) $(A^{-1})^{-1} = A$
- (c) AB is invertible and $(AB)^{-1} = B^{-1}A^{-1}$

Proof. (a) We find that:

$$(A^T)(A^{-1})^T = (A^{-1}A)^T = \mathbb{1}^T = \mathbb{1} \quad (45.3.24)$$

and similarly:

$$(A^{-1})^T(A^T) = (AA^{-1})^T = \mathbb{1}^T = \mathbb{1} \quad (45.3.25)$$

(b) We find that:

$$(A^{-1})(A) = \mathbb{1} = (A)(A^{-1}) \quad (45.3.26)$$

(c) We find that:

$$ABB^{-1}A^{-1} = AA^{-1} = \mathbb{1} \quad (45.3.27)$$

and

$$B^{-1}A^{-1}AB = B^{-1}B = \mathbb{1} \quad (45.3.28)$$

as desired. ■

Determinants

46.1 The determinant of a matrix

Definition (Determinant)

The determinant $\det : \mathbb{K}^n \rightarrow \mathbb{K}$ maps n vectors $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_N \in \mathbb{K}^n$ to a scalar $\det(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_N) \in \mathbb{K}$. It satisfies the following properties:

- (a) $\det(\dots, \alpha\mathbf{a} + \beta\mathbf{b}, \dots) = \alpha \det(\dots, \mathbf{a}, \dots) + \beta \det(\dots, \mathbf{b}, \dots)$.
- (b) $\det(\dots, \mathbf{a}, \dots, \mathbf{b}, \dots) = -\det(\mathbf{b}, \dots, \mathbf{a}, \dots)$
- (c) $\det(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n) = 1$

The determinant of a matrix \mathbf{A} with column vectors \mathbf{A}^i , $1 \leq i \leq n$ is the determinant of these column vectors:

$$\det \mathbf{A} \equiv \det(\mathbf{A}^1, \mathbf{A}^2, \dots, \mathbf{A}^n) \quad (46.1.1)$$

Note that:

$$\det(\dots, \mathbf{0}, \dots) = \det(\dots, \mathbf{v} - \mathbf{v}, \dots) = \det(\dots, \mathbf{v}, \dots) - \det(\dots, \mathbf{v}, \dots) = 0 \quad (46.1.2)$$

and:

$$\det(\dots, \mathbf{a}, \dots, \mathbf{a}, \dots) = -\det(\dots, \mathbf{a}, \dots, \mathbf{a}, \dots) \implies \det(\dots, \mathbf{a}, \dots, \mathbf{a}, \dots) = 0 \quad (46.1.3)$$

So the determinant of a matrix with a zero column is null, and so is the determinant of a matrix with a repeated column vector.

Also, we have that:

$$\det(\dots, \mathbf{a}, \dots, \alpha\mathbf{a}, \dots) = \alpha \det(\dots, \mathbf{a}, \dots, \mathbf{a}, \dots) = 0 \quad (46.1.4)$$

so the determinant of a matrix with two columns that are proportional to each other will also be zero. We can combine these results to state that the determinant of a matrix where one row is a linear combination of some of the others is also zero.

We summarize these results in the next theorem:

Proposition (Zero determinant matrices)

The following matrices have a zero determinant:

- (a) an entire row (or column) of zeros
- (b) a row (or column) that is a linear combination of other rows (or columns)

Proposition (Special determinants) The matrix of a diagonal matrix \mathbf{A} is given by the product of its diagonal elements.

The matrix of an upper or lower triangular matrix \mathbf{A} is also given by the product of its diagonal elements.

Proof. Consider a matrix $\mathbf{A} = \text{diag}(a_{11}, a_{22}, \dots, a_{nn})$. Then:

$$\det \mathbf{A} = \det(a_{11}\mathbf{e}_1, a_{22}\mathbf{e}_2, \dots, a_{nn}\mathbf{e}_n) = a_{11}a_{22}\dots a_{nn} \quad (46.1.5)$$

Instead, for an upper triangular matrix:

$$\det \mathbf{A} = \det\left(a_{11}\mathbf{e}_1, a_{12}\mathbf{e}_1 + a_{22}\mathbf{e}_2, \dots, \sum_i a_{in}\mathbf{e}_i\right) \quad (46.1.6)$$

$$= \det\left(a_{11}\mathbf{e}_1, a_{12}\mathbf{e}_1, \dots, \sum_i^0 a_{in}\mathbf{e}_i\right) + \det\left(a_{11}\mathbf{e}_1, a_{22}\mathbf{e}_2, \dots, \sum_i a_{in}\mathbf{e}_i\right) \quad (46.1.7)$$

$$= \det(a_{11}\mathbf{e}_1, a_{22}\mathbf{e}_2, \dots, a_{nn}\mathbf{e}_n) \quad (46.1.8)$$

$$= a_{11}a_{22}\dots a_{nn} \quad (46.1.9)$$

as desired. ■

Theorem (Determinant of matrix)

For a given $n \times n$ matrix \mathbf{A} , we have that:

$$\det \mathbf{A} = \sum_{\sigma \in S_n} \text{sgn}(\sigma) \prod_{k=1}^n a_{\sigma(i_k)k} \quad (46.1.10)$$

Proof. We start by writing:

$$\det \mathbf{A} = \det(\mathbf{A}^1, \mathbf{A}^2, \dots, \mathbf{A}^n) \quad (46.1.11)$$

$$= \det\left(\sum_{i_1} a_{i_1 1} \mathbf{e}_{i_1}, \sum_{i_2} a_{i_2 2} \mathbf{e}_{i_2}, \dots, \sum_k a_{i_n n} \mathbf{e}_{i_n}\right) \quad (46.1.12)$$

$$= \sum_{i_1} a_{i_1 1} \det\left(\mathbf{e}_{i_1}, \sum_{i_2} a_{i_2 2} \mathbf{e}_{i_2}, \dots, \sum_{i_n} a_{i_n n} \mathbf{e}_{i_n}\right) \quad (46.1.13)$$

$$= \sum_{i_1 i_2 \dots i_n} a_{i_1 1} a_{i_2 2} \dots a_{i_n n} \det(\mathbf{e}_{i_1}, \mathbf{e}_{i_2}, \dots, \mathbf{e}_{i_n}) \quad (46.1.14)$$

Now note that the only terms that survive out of this sum are $i_1 \neq i_2 \neq i_3 \neq \dots \neq i_n$. In other words, the only terms surviving are all the permutations of $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$, so we write:

$$\det \mathbf{A} = \sum_{\sigma \in S_n} \text{sgn}(\sigma) a_{\sigma(i_1) 1} a_{\sigma(i_2) 2} \dots a_{\sigma(i_n) n} \det(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n) \quad (46.1.15)$$

$$= \sum_{\sigma \in S_n} \text{sgn}(\sigma) \prod_{k=1}^n a_{\sigma(i_k) k} \quad (46.1.16)$$

as desired. ■

Proposition (Determinant properties)

Let's consider two $n \times n$ matrices \mathbf{A}, \mathbf{B} . Then:

- (a) $\det(\mathbf{A}^T) = \det \mathbf{A}$
- (b) $\det(\mathbf{AB}) = \det \mathbf{A} \cdot \det \mathbf{B}$
- (c) \mathbf{A} is bijective $\iff \det \mathbf{A} \neq 0$ and $\det \mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}}$

Proof. (a) Let a'_{ij} be the matrix elements of \mathbf{A}^T so that $a'_{ij} = a_{ji}$. Then:

$$\det \mathbf{A}^T = \sum_{\sigma \in S_n} \text{sgn}(\sigma) \prod_{k=1}^n a'_{\sigma(i_k) k} = \sum_{\sigma \in S_n} \text{sgn}(\sigma) \prod_{k=1}^n a_{k \sigma(i_k)} \quad (46.1.17)$$

$$= \sum_{\sigma \in S_n} \text{sgn}(\sigma^{-1}) \prod_{k=1}^n a_{\sigma^{-1}(i_k) k} = \sum_{\rho \in S_n} \text{sgn}(\rho^{-1}) \prod_{k=1}^n a_{\rho^{-1}(i_k) k} = \det \mathbf{A} \quad (46.1.18)$$

(b) Recall that $(\mathbf{AB})_{ik} = \sum_j a_{ij} b_{jk}$ implying that:

$$(\mathbf{AB})^k = \sum_{ij} a_{ij} b_{jk} \mathbf{e}_i = \sum_j b_{jk} \mathbf{A}^j \quad (46.1.19)$$

Hence, we get that:

$$\det AB = \det \left(\sum_{j_1} b_{j_1 1} \mathbf{A}^{j_1}, \sum_{j_2} b_{j_2 2} \mathbf{A}^{j_2}, \dots, \sum_{j_n} b_{j_n n} \mathbf{A}^{j_n} \right) \quad (46.1.20)$$

$$= \sum_{j_1, \dots, j_n} b_{j_1 1} b_{j_2 2} \dots b_{j_n n} \det(\mathbf{A}^{j_1}, \mathbf{A}^{j_2}, \dots, \mathbf{A}^{j_n}) \quad (46.1.21)$$

Again, we see that the only terms that survive are those where $j_1 \neq j_2 \neq \dots \neq j_n$, so we will get:

$$\det AB = \sum_{\sigma \in S_n} \text{sgn}(\sigma) b_{\sigma(j_1) 1} b_{\sigma(j_2) 2} \dots b_{\sigma(j_n) n} \det(\mathbf{A}^1, \mathbf{A}^2, \dots, \mathbf{A}^n) \quad (46.1.22)$$

$$= \det A \cdot \det B \quad (46.1.23)$$

as desired.

(c) (\implies) Suppose that A is bijective, and thus invertible. Then:

$$\det(AA^{-1}) = \det A \cdot \det A^{-1} = \det 1 = 1 \quad (46.1.24)$$

implying that $\det A \neq 0$.

(\Leftarrow) Suppose that A is not bijective, so that $\text{rk}(A) < n$. Therefore, there is at least one column vector, say \mathbf{A}^i , which may be written as a linear combination of some of the others:

$$\mathbf{A}^i = \sum_j \alpha_j \mathbf{A}^j \quad (46.1.25)$$

It then follows that

$$\det A = \det(\mathbf{A}^1, \dots, \mathbf{A}^i, \dots, \mathbf{A}^n) = \det(\mathbf{A}^1, \dots, \mathbf{A}^i, \dots, \mathbf{A}^n) \quad (46.1.26)$$

$$= \det(\mathbf{A}^1, \dots, \sum_j \alpha_j \mathbf{A}^j, \dots, \mathbf{A}^n) \quad (46.1.27)$$

$$= \sum_j \alpha_j \det(\mathbf{A}^1, \dots, \mathbf{A}^j, \dots, \mathbf{A}^j, \dots, \mathbf{A}^n) \quad (46.1.28)$$

$$= 0 \quad (46.1.29)$$

as desired.

Using (46.1.24):

$$\det A^{-1} = \frac{1}{\det A} \quad (46.1.30)$$

as desired. ■

46.2 Laplace expansion

Definition (Cofactor matrix)

Consider an $n \times n$ matrix \mathbf{A} , then we define the associated (i, j) matrix $\mathbf{A}(i, j)$ as the matrix \mathbf{A} with the i th row and j th column substituted with \mathbf{e}_i and \mathbf{e}_j^T respectively:

$$\mathbf{A}(i, j) = \begin{pmatrix} & & & \text{jth col} \\ & & 0 & \\ \mathbf{A} & \vdots & & \mathbf{A} \\ 0 \dots 0 & 1 & 0 \dots 0 \\ \mathbf{A} & \vdots & & \mathbf{A} \\ & 0 & & \end{pmatrix} \leftarrow i\text{th row} \quad (46.2.1)$$

The (i, j) cofactor coefficient is then defined as the determinant of $\mathbf{A}(i, j)$

$$C_{ij} = \det(\mathbf{A}(i, j)) \quad (46.2.2)$$

The cofactor matrix is the matrix \mathbf{C} whose elements are C_{ij} . The cofactor expansion in the i th row is defined as:

$$\text{cof}_i \mathbf{A} = \sum_k a_{ik} C_{ik} = a_{i1} C_{i1} + a_{i2} C_{i2} + \dots + a_{in} C_{in} \quad (46.2.3)$$

It turns out that the cofactor matrix is especially important in evaluating the inverse of matrices. It is therefore important to be able to calculate the determinant of \mathbf{A}_{ij} more easily.

Proposition (Cofactor matrix calculation)

Let \mathbf{A} be a $n \times n$ matrix and let $\tilde{\mathbf{A}}_{ij}$ be the matrix \mathbf{A} with the i th row and j th column removed. Then:

$$C_{ij} = (-1)^{i+j} \det(\tilde{\mathbf{A}}(i, j)) \quad (46.2.4)$$

Proof. Note that the determinant only acquires a sign change when moving columns, and the same goes for rows since $\det(\mathbf{A}^T) = \det \mathbf{A}$. Hence, we may move the i th row and j th column to the first row and column respectively. To do so we must perform $i - 1$ row exchanges followed by $j - 1$ exchanges.¹ If we define $\tilde{\mathbf{A}}_{ij}$ to be the matrix with the i th

¹we can't just exchange the i th row with the first row, since this would alter the order of the rows, and would not give the matrix \mathbf{A} with the i th row removed.

and j th rows removed, then if:

$$\mathbf{B}(i, j) \equiv \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \tilde{\mathbf{A}}(i, j) & \\ 0 & & & \end{pmatrix} \quad (46.2.5)$$

$$\implies \det(\mathbf{B}(i, j)) = (-1)^{i+j+2} \det(\mathbf{A}(i, j)) = \det(\tilde{\mathbf{A}}(i, j)) \quad (46.2.6)$$

so that:

$$\det(\mathbf{A}(i, j)) = (-1)^{i+j} \det(\tilde{\mathbf{A}}(i, j)) \quad (46.2.7)$$

■

This is a much easier formula to use when evaluating the cofactor matrix, since instead of evaluating the determinant of an $n \times n$ matrix, we're evaluating the determinant of an $(n - 1) \times (n - 1)$ matrix.

Theorem (Laplace expansion)

For a given $n \times n$ matrix \mathbf{A} with cofactor matrix \mathbf{C} :

$$(\det \mathbf{A}) \mathbb{1} = \mathbf{C}^T \mathbf{A} \quad (46.2.8)$$

so that:

$$\det \mathbf{A} = \sum_k (-1)^{k+i} A_{kj} \det(\tilde{\mathbf{A}}(k, i)) \quad (46.2.9)$$

Proof. We find that:

$$(\mathbf{C}^T \mathbf{A})_{ij} = \sum_k C_{ki} A_{kj} \quad (46.2.10)$$

$$= \sum_k \det(\mathbf{A}(k, i)) A_{kj} \quad (46.2.11)$$

$$= \sum_k A_{kj} \det(\mathbf{A}^1 - A_{k1} \mathbf{e}_k, \mathbf{A}^2 - A_{k2} \mathbf{e}_k, \dots, \mathbf{e}_k, \dots, \mathbf{A}^n - A_{kn} \mathbf{e}_n) \quad (46.2.12)$$

Now note that

$$\det(\mathbf{A}^1 - A_{k1} \mathbf{e}_k, \dots, \mathbf{e}_k, \dots) = \det(\mathbf{A}^1, \dots, \mathbf{e}_k, \dots) - A_{k1} \det(\mathbf{e}_k, \dots, \mathbf{e}_k, \dots) \quad (46.2.13)$$

$$= \det(\mathbf{A}^1, \dots, \mathbf{e}_k, \dots) \quad (46.2.14)$$

Repeating this process we find that:

$$(\mathbf{C}^T \mathbf{A})_{ij} = \sum_k A_{kj} \det(\mathbf{A}^1, \mathbf{A}^2, \dots, \mathbf{e}_k, \dots, \mathbf{A}^n) \quad (46.2.15)$$

$$= \sum_k \det(\mathbf{A}^1, \mathbf{A}^2, \dots, \mathbf{A}^j, \dots, \mathbf{A}^n) \quad (46.2.16)$$

$$= \det(\mathbf{A}^1, \mathbf{A}^2, \dots, \mathbf{A}^i, \dots, \mathbf{A}^n) \delta_{ij} \quad (46.2.17)$$

$$= \det \mathbf{A} \delta_{ij} \quad (46.2.18)$$

implying that:

$$\mathbf{C}^T \mathbf{A} = \det \mathbf{A} \mathbf{1} \quad (46.2.19)$$

as desired. ■

Example. Consider a 3×3 matrix:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \quad (46.2.20)$$

We calculate the cofactor expansion in the first column: Then we find that:

$$C_{11} = a_{22}a_{33} - a_{23}a_{32}, \quad C_{21} = a_{13}a_{32} - a_{12}a_{23}, \quad C_{31} = a_{12}a_{23}a_{22}a_{13} \quad (46.2.21)$$

so:

$$\det \mathbf{A} = C_{11}a_{11} + C_{21}a_{21} + C_{31}a_{31} \quad (46.2.22)$$

$$= a_{11}(a_{22}a_{33} - a_{23}a_{32}) + a_{21}(a_{13}a_{32} - a_{12}a_{23}) + a_{31}(a_{12}a_{23}a_{22}a_{13}) \quad (46.2.23)$$

Note that had we used the cofactor expansion in any other column (or also row), the d

◀

Proposition (Cofactor orthogonality with rows)

We have that the j th row \mathbf{A}^j of a matrix \mathbf{A} is orthogonal to the i th row of its cofactor matrix \mathbf{C} , for $i \neq j$. So:

$$\mathbf{A}^j \cdot \mathbf{C}_i = \sum_k a_{jk} C_{ik} = 0 \quad (46.2.24)$$

Proof. For $i \neq j$, we have that

$$\mathbf{A}^j \cdot \mathbf{C}_i = \sum_k a_{jk} C_{ik} = a_{j1}C_{i1} + a_{j2}C_{i2} + \dots + a_{jn}C_{in} \quad (46.2.25)$$

Our goal is to find some other matrix B whose cofactor expansion is like this. To find the form of this matrix, we note that:

$$a_{j1}C_{i1} + a_{j2}C_{i2} + \dots + a_{jn}C_{in} = b_{i1}C_{i1} + b_{i2}C_{i2} + \dots + b_{in}C_{in} \quad (46.2.26)$$

Firstly, since C_{ij} takes all the entries of A without the i th row and j th column, it follows that for these cofactors to coincide with those of B , all the elements of B except for the i th row are identical to those of A .

The only change is that for $a_{jk} = b_{ik}$, $k = 1, 2, \dots, n$, so the i th row of B is the j th row of A . However, note that the j th row of B must also be the j th row of A as we argued in the previous paragraph, so B has two repeated rows. Therefore, its determinant/cofactor expansion must vanish:

$$a_{j1}C_{i1} + a_{j2}C_{i2} + \dots + a_{jn}C_{in} = b_{i1}C_{i1} + b_{i2}C_{i2} + \dots + b_{in}C_{in} = 0 \quad (46.2.27)$$

■

Definition (Adjoint matrix)

The adjoint $\text{adj } A$ of a matrix A is the transpose of its cofactor matrix C :

$$\text{adj } A = C^T \implies (\det A)\mathbb{1} = (\text{adj } A)A \quad (46.2.28)$$

Theorem (Inverse of matrix)

Let A be an invertible $n \times n$ matrix, then:

$$A^{-1} = \frac{1}{\det A} \text{adj } A \quad (46.2.29)$$

Proof. We find that:

$$(\det A)\mathbb{1} = C^T A \implies (\det A)A^{-1} = \text{adj } A \implies A^{-1} = \frac{1}{\det A} \text{adj } A \quad (46.2.30)$$

as desired. ■

Example. Consider the 2×2 matrix:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad ad - bc \neq 0 \quad (46.2.31)$$

We find that:

$$\det A = ad - bc \quad (46.2.32)$$

Furthermore:

$$C_{11} = d, \quad C_{12} = -c, \quad C_{21} = -b, \quad C_{22} = d \quad (46.2.33)$$

giving the cofactor matrix:

$$C = \begin{pmatrix} d & -c \\ -b & d \end{pmatrix} \quad (46.2.34)$$

Consequently

$$\text{adj } A = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \implies A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \quad (46.2.35)$$

◀

46.3 Cramer's rule

With our newfound knowledge of determinants, we are now ready to formulate yet another method to solve linear systems. Previously we have discussed Gauss-Jordan elimination, as well as inverting matrices as methods of solutions.

Theorem (Cramer's rule)

Consider a linear system of equations $Ax = b$. Let B_i be the matrix A with the i th column replaced with b . Then the solution to the system is given by:

$$x_i \det A = \det B_i \quad (46.3.1)$$

Proof. We find that:

$$\det B_i = \det(A^1, \dots, b, \dots, A^n) \quad (46.3.2)$$

$$= \sum_i b_i \det(A^1, \dots, e_i, \dots, A^n) \quad (46.3.3)$$

$$= \sum_i A_{ij} x_j \det(A^1, \dots, e_i, \dots, A^n) \quad (46.3.4)$$

$$= \sum_i x_j \det(A^1, \dots, A^j, \dots, A^n) \quad (46.3.5)$$

$$= x_i \det(A^1, \dots, A^i, \dots, A^n) = x_i \det A \quad (46.3.6)$$

as desired. Hence, if A is invertible then the solutions are given by:

$$x_i = \frac{\det B_i}{\det A} \quad (46.3.7)$$

■

Example. Consider the following system:

$$\begin{cases} x + 2y + 3z = 0 \\ 2x + 3y + 4z = 1 \\ 3x + 4y + 6z = 2 \end{cases} \quad (46.3.8)$$

Then we see that:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 6 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} \quad (46.3.9)$$

so that:

$$\mathbf{B}_1 = \begin{pmatrix} 0 & 2 & 3 \\ 1 & 3 & 4 \\ 2 & 4 & 6 \end{pmatrix}, \quad \mathbf{B}_2 = \begin{pmatrix} 1 & 0 & 3 \\ 2 & 1 & 4 \\ 3 & 2 & 6 \end{pmatrix}, \quad \mathbf{B}_3 = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 3 & 1 \\ 3 & 4 & 2 \end{pmatrix} \quad (46.3.10)$$

Now we evaluate the determinants using the Laplace expansion:

$$\det \mathbf{A} = (18 - 16) - 2(12 - 12) + 3(8 - 9) = -1 \quad (46.3.11)$$

$$\det \mathbf{B}_1 = -2(6 - 8) + 3(4 - 6) = -2 \quad (46.3.12)$$

$$\det \mathbf{B}_2 = (6 - 8) + 3(4 - 3) = 1 \quad (46.3.13)$$

$$\det \mathbf{B}_3 = (6 - 4) - 2(4 - 3) = 0 \quad (46.3.14)$$

so that:

$$x = 2, \quad y = -1, \quad z = 0 \quad (46.3.15)$$



Inner product spaces

47.1 Inner products

Definition (Inner products)

An inner product on a vector space V defined over $\mathbb{K} = \mathbb{R}$ (or \mathbb{C}) is a map:

$$\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{K} \quad (47.1.1)$$

satisfying:

- S1. $\langle \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{w}, \mathbf{v} \rangle$ for a symmetric scalar product
(or $\langle \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{w}, \mathbf{v} \rangle^*$ for hermitian inner product)
- S2. $\langle \mathbf{v}, \alpha \mathbf{u} + \beta \mathbf{w} \rangle = \alpha \langle \mathbf{v}, \mathbf{u} \rangle + \beta \langle \mathbf{v}, \mathbf{w} \rangle$
- S3. $\langle \mathbf{v}, \mathbf{v} \rangle \geq 0$ with equality holding only for $\mathbf{v} = 0$
for all $\mathbf{v}, \mathbf{w}, \mathbf{u} \in V, \alpha, \beta \in \mathbb{K}$.

Example.

- i. Minkowski product: for $\mathbf{v} = v^\mu \mathbf{e}_\mu \in \mathbb{R}^4$ and $\mathbf{w} = w^\mu \mathbf{e}_\mu \in \mathbb{R}^4$, the Minkowski product is defined as:

$$\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T \boldsymbol{\eta} \mathbf{v} \quad (47.1.2)$$

where $\boldsymbol{\eta} = \text{diag}(1, -1, -1, -1)$.

- ii. Functional products: for $f, g \in C^0([a, b])$, we define their inner product as:

$$\langle f, g \rangle = \int_a^b f(x)^* g(x) dx \quad (47.1.3)$$

Proposition (Inner product is well-defined)

Let $f, g \in \text{End}(V)$ be two endomorphisms on V , then:

$$\langle \mathbf{v}, f(\mathbf{w}) \rangle = \langle \mathbf{v}, g(\mathbf{w}) \rangle, \forall \mathbf{v}, \mathbf{w} \in V \implies f = g \quad (47.1.4)$$

Proof. Let $\mathcal{B} = \{\mathbf{e}_i\}$ be an orthonormal basis for V . It follows that for any $\mathbf{v}, \mathbf{w} \in V$ (we assume a hermitian inner product, the proof for a symmetric scalar product is similar):

$$\langle \mathbf{v}, f(\mathbf{w}) \rangle = \langle \mathbf{v}, g(\mathbf{w}) \rangle \iff \langle \mathbf{v}, f(\mathbf{w}) - g(\mathbf{w}) \rangle = 0 \quad (47.1.5)$$

$$\iff \sum_{ij} v_i^* w_j \langle \mathbf{e}_i, f(\mathbf{e}_j) - g(\mathbf{e}_j) \rangle = 0 \quad (47.1.6)$$

Since this applies for any \mathbf{v}, \mathbf{w} , we must therefore have that:

$$\langle \mathbf{e}_i, h(\mathbf{e}_j) \rangle \equiv \langle \mathbf{e}_i, f(\mathbf{e}_j) - g(\mathbf{e}_j) \rangle = 0, \forall i, j = 1, 2, \dots \quad (47.1.7)$$

where we defined $h = f - g \in \text{End}(V)$. Consequently, $h(\mathbf{e}_j)$ cannot have any component along any of the basis vectors for V , and must therefore be zero. Hence $f = g$ as desired. ■

Definition (Orthogonal complement) For a subspace W of V , we define the orthogonal complement W^\perp as:

$$W^\perp = \{\mathbf{v} \in V : \langle \mathbf{v}, \mathbf{w} \rangle = 0, \forall \mathbf{w} \in W\} \quad (47.1.8)$$

47.2 Projectors

Proposition (Properties of orthogonal complements) For a subspace W of V with orthogonal complement W^\perp , we have that:

- (i) $W^\perp \subset V$
- (ii) $W \cap W^\perp = \{\mathbf{0}\}$
- (iii) $\dim W + \dim W^\perp = \dim V$

Proof. (i) Trivial.

(ii),(iii) Firstly note that $W \oplus W^\perp = V$. Indeed, given a vector $\mathbf{v} \in V$, we can decompose it as:

$$\mathbf{v} = \underbrace{\langle \mathbf{w}, \mathbf{v} \rangle \mathbf{w}}_{\in W} + \underbrace{(\mathbf{v} - \langle \mathbf{w}, \mathbf{v} \rangle \mathbf{w})}_{\in W^\perp} \quad (47.2.1)$$

Therefore, we must have that $\dim W + \dim W^\perp = \dim V$, as well as $W \cap W^\perp = \{\mathbf{0}\}$ from the properties of direct sums. ■

Definition ((Orthogonal) Projection operators) Let $V = U \oplus W$. We define a projection from V to W as a map Π satisfying:

$$\Pi : V \rightarrow W \quad (47.2.2)$$

$$\mathbf{u} + \mathbf{w} \rightarrow \mathbf{w} \quad (47.2.3)$$

with $\mathbf{u} + \mathbf{w}$. Clearly, $\Pi^2 = \Pi$ so it is idempotent.

If $U = V^\perp$ then we say that Π is an orthogonal projection operator.

Theorem (All idempotent maps are projective)

All idempotent maps are projections.

Proof. Let Π be an idempotent map so that $\Pi^2 = \Pi$. Let \mathbf{v} , then clearly we have that:

$$\mathbf{v} = \Pi(\mathbf{v}) + (\mathbf{v} - \Pi(\mathbf{v})) = \Pi(\mathbf{v}) + (\mathbb{1} - \Pi)(\mathbf{v}) \quad (47.2.4)$$

Now let us define $W = \{\Pi(\mathbf{v}) \mid \forall \mathbf{v} \in V\} = \text{Im}(\Pi)$ and $U = \{\mathbf{v} - \Pi(\mathbf{v}) : \forall \mathbf{v} \in V\} = \text{Im}(\mathbb{1} - \Pi)$. Since these are both images of linear transformations, we have that U, W are subspaces of V . Note also that $\Pi(\mathbf{w}) = 0$ for all $\mathbf{w} \in W$ and $\Pi(\mathbf{u}) = \mathbf{u}$. Consequently $V = U \otimes W$, with:

$$\Pi(\mathbf{v}) = \Pi(\mathbf{u} + \mathbf{w}) = \Pi(\mathbf{u}) + \Pi(\mathbf{w}) = \mathbf{u} \quad (47.2.5)$$

proving that Π is indeed a projector. ■

47.3 Inner products and matrices

Definition (Adjoint, Hermitian and Unitary linear map)

For a linear map $f \in \text{End}(V)$ on V defined over \mathbb{K} , its adjoint map $f^\dagger \in \text{End}(V)$ is defined so that it satisfies:

$$\langle \mathbf{v}, f(\mathbf{w}) \rangle = \langle f^\dagger(\mathbf{v}), \mathbf{w} \rangle, \quad \forall \mathbf{v}, \mathbf{w} \in V \quad (47.3.1)$$

If $\mathbb{K} = \mathbb{R}$ (or \mathbb{C}), then a symmetric (or hermitian) linear map f satisfies $f = f^\dagger$, while an orthogonal (or unitary) linear map U satisfies $U^{-1} = U^\dagger$.

Proposition (Adjoint map properties)

For a linear map $f \in \text{End}(V)$, the following must hold:

- (i) f^\dagger is unique
- (ii) $(f^\dagger)^\dagger = f$
- (iii) $(f \circ g)^\dagger = g^\dagger \circ f^\dagger$
- (iv) $(f^{-1})^\dagger = (f^\dagger)^{-1}$

- (i) Let g, h be adjoint maps of f . Then $\langle \mathbf{v}, f(\mathbf{w}) \rangle = \langle g(\mathbf{v}), \mathbf{w} \rangle = \langle \mathbf{v}, f(\mathbf{w}) \rangle = \langle h(\mathbf{v}), \mathbf{w} \rangle$ which by previous proposition implies that $g = h$.
- (ii) For all $\mathbf{v}, \mathbf{w} \in V$, we have that $\langle \mathbf{v}, f(\mathbf{w}) \rangle = \langle f^\dagger(\mathbf{v}), \mathbf{w} \rangle = \langle \mathbf{v}, (f^\dagger)^\dagger(\mathbf{w}) \rangle$ which be the same proposition as before implies that $(f^\dagger)^\dagger = f$
- (iii) $\langle \mathbf{v}, (f \circ g)^\dagger(\mathbf{w}) \rangle = \langle f(g(\mathbf{v})), \mathbf{w} \rangle = \langle g(\mathbf{v}), f^\dagger(\mathbf{w}) \rangle = \langle \mathbf{v}, (g^\dagger \circ f^\dagger)(\mathbf{w}) \rangle$.
- (iv) $f \circ f^{-1} = \text{id}_V \implies (f^{-1})^\dagger \circ f^\dagger = \text{id}_V \implies (f^{-1})^\dagger = (f^\dagger)^{-1}$ where we used (iii) to take the adjoint of both sides in the first implication.

Note that if we have an orthonormal basis $\mathcal{B} = \{\mathbf{e}_i\}$ for V , then given any $f \in \text{End}(V)$, we have that:

$$f(\mathbf{e}_i) = \sum_j A_{ij} \mathbf{e}_j \implies A_{ij} = \langle \mathbf{e}_j, f(\mathbf{e}_i) \rangle \quad (47.3.2)$$

so we can use inner products to find the matrix elements of linear maps. It is easy to see that if $f \in \text{End}(V)$ defined over \mathbb{C} has matrix elements A_{ij} in a given basis, then its adjoint f^\dagger will have matrix elements B_{ij} in the same basis given by:

$$B_{ij} = \langle \mathbf{e}_i, f^\dagger(\mathbf{e}_j) \rangle = \langle f(\mathbf{e}_i), \mathbf{e}_j \rangle = \langle \mathbf{e}_j, f(\mathbf{e}_i) \rangle^* = A_{ji}^* \quad (47.3.3)$$

Hence the matrices $\mathbf{A}, \mathbf{A}^\dagger$ representing f, f^\dagger respectively satisfy $\mathbf{A}^\dagger = (\mathbf{A}^*)^T$.

It follows that Hermitian maps have matrix representations $\mathbf{A} = \mathbf{A}^T = (\mathbf{A}^*)^T$ and Unitary maps have matrix representations $\mathbf{A}^{-1} = \mathbf{A}^T = (\mathbf{A}^*)^T$.

Proposition (Alternative definition of Unitarity) Let $U \in \text{End}(V)$ is a unitary map, then $\langle U(\mathbf{v}), U(\mathbf{w}) \rangle = \langle \mathbf{v}, \mathbf{w} \rangle$ for all $\mathbf{v}, \mathbf{w} \in V$ is an equivalent definition.

Proof. We find $\langle U(\mathbf{v}), U(\mathbf{w}) \rangle = \langle \mathbf{v}, (U^\dagger \circ U)(\mathbf{w}) \rangle = \langle \mathbf{v}, \mathbf{w} \rangle$. Hence by the well-definedness of inner products, $U^\dagger \circ U = \text{id}_V \iff U^{-1} = U^\dagger$. ■

47.4 Bilinear and Sesquilinear forms

In the previous section we looked at properties of inner product spaces over real or complex vector spaces. It turns out that when we remove the condition for the inner product to be positive semi-definite we get some interesting new forms.

Definition (Bilinear/Sesquilinear form) A Bilinear form on a vector space V defined over \mathbb{R} is a map $T : V \times V \rightarrow \mathbb{R}$ linear in both of its terms:

$$T(\alpha \mathbf{v} + \beta \mathbf{w}, \mathbf{u}) = \alpha T(\mathbf{v}, \mathbf{u}) + \beta T(\mathbf{w}, \mathbf{u}) \quad (47.4.1)$$

$$T(\mathbf{u}, \alpha \mathbf{v} + \beta \mathbf{w}) = \alpha T(\mathbf{u}, \mathbf{v}) + \beta T(\mathbf{u}, \mathbf{w}) \quad (47.4.2)$$

A bilinear form is symmetric if $T(\mathbf{v}, \mathbf{w}) = T(\mathbf{w}, \mathbf{v})$.

A Sesquilinear form on a vector space V defined over \mathbb{C} is a map $T : V \times V \rightarrow \mathbb{C}$ linear in both of its terms:

$$T(\alpha\mathbf{v} + \beta\mathbf{w}, \mathbf{u}) = \alpha^*T(\mathbf{v}, \mathbf{u}) + \beta^*T(\mathbf{w}, \mathbf{u}) \quad (47.4.3)$$

$$T(\mathbf{u}, \alpha\mathbf{v} + \beta\mathbf{w}) = \alpha T(\mathbf{u}, \mathbf{v}) + \beta T(\mathbf{u}, \mathbf{w}) \quad (47.4.4)$$

A sesquilinear form is hermitian if $T(\mathbf{v}, \mathbf{w}) = T(\mathbf{w}, \mathbf{v})$.

Example. An important example of a Bilinear form often used in Special relativity is:

$$T(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{A} \mathbf{y} \quad (47.4.5)$$

where $\mathbf{x}, \mathbf{y} \in V$ and $\mathbf{A} \in \text{Mat}_n(V)$. Indeed, we prove linearity in the first argument as follows:

$$T(\alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z}) = (\alpha\mathbf{x} + \beta\mathbf{y})^T \mathbf{A} \mathbf{z} \quad (47.4.6)$$

$$= (\alpha\mathbf{x}^T + \beta\mathbf{y}^T) \mathbf{A} \mathbf{z} \quad (47.4.7)$$

$$= \alpha\mathbf{x}^T \mathbf{A} \mathbf{z} + \beta\mathbf{y}^T \mathbf{A} \mathbf{z} \quad (47.4.8)$$

$$= \alpha T(\mathbf{x}, \mathbf{z}) + \beta T(\mathbf{y}, \mathbf{z}) \quad (47.4.9)$$

The proof is similar for the linearity in second argument. We can extend this example to sesquilinear forms by defining:

$$T(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\dagger \mathbf{A} \mathbf{y} \quad (47.4.10)$$

where $\mathbf{x}, \mathbf{y} \in V$ and $\mathbf{A} \in \text{Mat}_n(V)$. Here \dagger denotes conjugate transposition. We again prove linearity in the first argument as follows:

$$T(\alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z}) = (\alpha\mathbf{x} + \beta\mathbf{y})^\dagger \mathbf{A} \mathbf{z} \quad (47.4.11)$$

$$= (\alpha\mathbf{x}^\dagger + \beta\mathbf{y}^\dagger) \mathbf{A} \mathbf{z} \quad (47.4.12)$$

$$= \alpha\mathbf{x}^\dagger \mathbf{A} \mathbf{z} + \beta\mathbf{y}^\dagger \mathbf{A} \mathbf{z} \quad (47.4.13)$$

$$= \alpha T(\mathbf{x}, \mathbf{z}) + \beta T(\mathbf{y}, \mathbf{z}) \quad (47.4.14)$$

Note also that if \mathbf{A} is a symmetric matrix then:

$$T(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{A}) \mathbf{y} = (\mathbf{A}^T \mathbf{x})^T \mathbf{y} = \mathbf{y}^T \mathbf{A} \mathbf{x} \quad (47.4.15)$$

so T is a symmetric bilinear form. We can extend this result to sesquilinear forms quite easily by letting \mathbf{A} be hermitian. Then

$$T(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^\dagger \mathbf{A}) \mathbf{y} = (\mathbf{A}^\dagger \mathbf{x})^\dagger \mathbf{y} = \mathbf{y}^\dagger \mathbf{A} \mathbf{x} \quad (47.4.16)$$



It turns out that all bilinear/sesquilinear forms may be expressed in the form of the previous example. Indeed, let T be a (bilinear) sesquilinear form on a (real) complex vector space V . Let $\{\mathbf{e}_i\}$ be an ordered basis of V , then:

$$T(\mathbf{x}, \mathbf{y}) = \sum_{ij} x_i^* y_j T(\mathbf{e}_i, \mathbf{e}_j) \quad (47.4.17)$$

If we let $A_{ij} = T(\mathbf{e}_i, \mathbf{e}_j)$ then clearly:

$$\mathbf{x}^\dagger \mathbf{A} \mathbf{y} = \sum_{ij} x_i^* A_{ij} y_j = \sum_{ij} x_i^* y_j T(\mathbf{e}_i, \mathbf{e}_j) = T(\mathbf{x}, \mathbf{y}) \quad (47.4.18)$$

as desired.

Theorem (Matrix representation of forms) A bilinear/sesquilinear form T over a real/complex vector space V has an associated matrix representation in a given basis $\{\mathbf{e}_i\}$ of V :

$$\mathbf{A} = \begin{pmatrix} T(\mathbf{e}_1, \mathbf{e}_1) & T(\mathbf{e}_1, \mathbf{e}_2) & \dots \\ T(\mathbf{e}_2, \mathbf{e}_1) & T(\mathbf{e}_2, \mathbf{e}_2) & \dots \\ \vdots & \ddots & \vdots \end{pmatrix} \quad (47.4.19)$$

Analogously to linear maps, one can also perform changes of basis for sesquilinear/bilinear forms. Suppose that in the basis $\{\mathbf{e}_i\}$ the form T is represented by \mathbf{A} so that

$$A_{ij} = T(\mathbf{e}_i, \mathbf{e}_j) \quad (47.4.20)$$

Let us introduce a new basis $\{\mathbf{e}'_i\}$ such that:

$$\mathbf{e}'_i = \sum_m c_{mi} \mathbf{e}_m \quad (47.4.21)$$

then we find

$$A'_{ij} = T(\mathbf{e}'_i, \mathbf{e}'_j) = \sum_{mn} c_{mi}^* c_{nj} T(\mathbf{e}_m, \mathbf{e}_n) = \sum_{mn} c_{mi}^* A_{mn} c_{nj} \quad (47.4.22)$$

Consequently, if we define a change of basis matrix \mathbf{P} with components $P_{mn} = c_{mn}$ then we get:

$$\mathbf{A}' = \mathbf{P}^\dagger \mathbf{A} \mathbf{P} \quad (47.4.23)$$

We interpret this result as usual. \mathbf{P} converts our vector from the original unprimed basis to the new primed basis:

$$\mathbf{x} = \sum_i x'_i \mathbf{e}'_i = \sum_{ij} x'_i c_{ji} \mathbf{e}_j = \sum_j x_j \mathbf{e}_j \quad (47.4.24)$$

where

$$x'_j = \sum_i x'_i c_{ji} \iff [\mathbf{x}]' = \mathbf{P}[\mathbf{x}] \quad (47.4.25)$$

Therefore if we want to calculate the form $\mathbf{x}^\dagger \mathbf{A} \mathbf{y}$ then we need a P^\dagger to the left of \mathbf{A} to convert the components of \mathbf{x}^\dagger to the primed basis, and a P to the right of \mathbf{A} to convert the components of \mathbf{y} .

Eigen-everything

48.1 Finding eigenvalues and eigenvectors

Definition (Eigenvalue and eigenvector)

Let $f : V \rightarrow V$ be a linear map on V over \mathbb{F} . We say that $\lambda \in \mathbb{F}$ is an eigenvalue of f if $\exists \mathbf{v} \in V$, s.t. $\mathbf{v} \neq \mathbf{0}$, known as an eigenvector, such that:

$$f(\mathbf{v}) = \lambda \mathbf{v} \quad (48.1.1)$$

The eigenspace of λ is defined as:

$$\text{Eig}_f(\lambda) \equiv \text{Ker}(f - \lambda \text{id}_V) \subseteq V \quad (48.1.2)$$

For a map f to have non-trivial eigenvalues, we require that:

$$\dim \text{Eig}_f(\lambda) = \text{Ker}(f - \lambda \text{id}_V) > 0 \implies \text{rk}(f - \lambda \text{id}_V) < \dim V \quad (48.1.3)$$

This is equivalent, by the proposition on linear systems and invertibility, to setting

$$\det(f - \lambda \text{id}_V) = 0 \quad (48.1.4)$$

We could have also seen this by noting that if $f - \lambda \text{id}_V$ were invertible, then there would only be one \mathbf{v} in $\text{Ker}(f - \lambda \text{id}_V)$, which must be $\mathbf{0}$. Therefore $f - \lambda \text{id}_V$ cannot be invertible, yielding (48.1.4).

Definition (Characteristic polynomial)

The characteristic polynomial of a map $f : V \rightarrow V$ is defined as:

$$\chi_f(\lambda) = \det(f - \lambda \text{id}_V) \quad (48.1.5)$$

To find the eigenvalues of a matrix, it suffices to:

- (i) Compute its characteristic polynomial.
- (ii) Find the roots λ of $\chi_f(\lambda)$.

(iii) For each solution λ , find the corresponding eigenspace by solving:

$$(f - \lambda \text{id}_V) \mathbf{v} = 0 \quad (48.1.6)$$

using one of the methods introduced for solving linear systems.

Example. Let us find the eigenvalues and eigenspaces of

$$\mathbf{A} = \begin{pmatrix} 4 & 0 & 4 \\ 0 & 4 & 4 \\ 4 & 4 & 8 \end{pmatrix} \quad (48.1.7)$$

Its characteristic polynomial is:

$$\chi_{\mathbf{A}}(\lambda) = \begin{vmatrix} 4 - \lambda & 0 & 4 \\ 0 & 4 - \lambda & 4 \\ 4 & 4 & 8 - \lambda \end{vmatrix} \quad (48.1.8)$$

$$= (4 - \lambda)((4 - \lambda)(8 - \lambda) - 16) - 16(4 - \lambda) \quad (48.1.9)$$

$$= (4 - \lambda)(\lambda^2 - 12\lambda) \quad (48.1.10)$$

$$= \lambda(4 - \lambda)(\lambda - 12) \quad (48.1.11)$$

Clearly, the solutions to $\chi_{\mathbf{A}}(\lambda) = 0$ are $\lambda = 0, 4, 12$.

For $\lambda_1 = 0$, we need:

$$\begin{pmatrix} 4 & 0 & 4 \\ 0 & 4 & 4 \\ 4 & 4 & 8 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \implies x = -z, y = -z \quad (48.1.12)$$

giving an eigenspace:

$$\text{Eig}_{\mathbf{A}}(\lambda_1) = \left\{ k \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}, \forall k \in \mathbb{R}^* \right\} \quad (48.1.13)$$

Similarly, for $\lambda_1 = 4$, we need:

$$\begin{pmatrix} 0 & 0 & 4 \\ 0 & 0 & 4 \\ 4 & 4 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \implies x = -y, z = 0 \quad (48.1.14)$$

giving an eigenspace:

$$\text{Eig}_{\mathbf{A}}(\lambda_1) = \left\{ k \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \forall k \in \mathbb{R}^* \right\} \quad (48.1.15)$$

Finally, for $\lambda_1 = 12$, we need:

$$\begin{pmatrix} -8 & 0 & 4 \\ 0 & -8 & 4 \\ 4 & 4 & -4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \implies z = 2x, y = x \quad (48.1.16)$$

giving an eigenspace:

$$\text{Eig}_A(\lambda_1) = \left\{ k \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}, \forall k \in \mathbb{R}^* \right\} \quad (48.1.17)$$

◀

Proposition (*Characteristic polynomial*)

For a matrix $A \in \text{Mat}_n(\mathbb{K})$ with characteristic polynomial $\chi_A(\lambda) = \sum_{i=0}^n c_i \lambda^i$, the following hold:

- (i) $\chi_{PAP^{-1}} = \chi_A$ for all $P \in \text{Mat}_n(\mathbb{K})$
- (ii) $c_n = (-1)^n, c_{n-1} = (-1)^{n-1} \text{tr} A, c_0 = \det A$

Proof. (i) We have that:

$$\det(PAP^{-1} - \lambda \mathbb{1}) = \det(P(A - \lambda \mathbb{1})P^{-1}) = \det(A - \lambda \mathbb{1}) \quad (48.1.18)$$

as desired.

(ii) We have that:

$$c_0 = \chi_A(0) = \det A \quad (48.1.19)$$

Furthermore

$$\chi_A(\lambda) = \prod_{i=1}^n (A_{ii} - \lambda) + o(\lambda^{n-2}) \quad (48.1.20)$$

$$= (-1)^n \lambda^n + (-1)^{n-1} \sum_{i=1}^n A_{ii} + o(\lambda^{n-1}) \quad (48.1.21)$$

$$= (-1)^n \lambda^n + (-1)^{n-1} \text{tr} A + o(\lambda^{n-1}) \quad (48.1.22)$$

implying that $c_n = (-1)^n$ and $c_{n-1} = (-1)^{n-1} \text{tr} A$.

■

48.2 Matrix diagonalization

Definition (*Diagonalized map*)

A linear map $f : V \rightarrow V$ can be diagonalized *iff* there exists a basis of V which makes the matrix representation of f diagonal, that is, if it is similar to a diagonal matrix.

Theorem (*Diagonalizability*)

A linear map $f : V \rightarrow V$ can be diagonalised *iff* there exists a basis of V consisting of eigenvectors of f . In this basis, the matrix representation of f is $\text{diag}(\lambda_i)$ where λ_i are the eigenvalues of f .

Proof. (\implies) Suppose that f can be diagonalized into the form $\text{diag}(c_i)$ in some basis $\{\mathbf{v}_i\}$. This implies that:

$$f(\mathbf{v}_i) = \sum_j A_{ji} \mathbf{v}_j = \sum_j \delta_{ji} c_i \mathbf{v}_i \quad (48.2.1)$$

which gives $f(\mathbf{v}_i) = c_i \mathbf{v}_i$, as desired.

(\impliedby) Let $\mathbf{P} = (\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_n)$ be the transition matrix from some basis \mathcal{B} to the set $\{\mathbf{v}_i\}$ of eigenvectors (with respective eigenvalues λ_i). Of course, to perform the required change of basis we need the set of eigenvectors $\{\mathbf{v}_i\}$ to form a basis of V . Then, we find that the matrix representation \mathbf{A} of f in \mathcal{B} :

$$\mathbf{A}' = \mathbf{P}^{-1} \mathbf{A} \mathbf{P} = \mathbf{P}^{-1} \mathbf{A} (\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_n) \quad (48.2.2)$$

$$= \mathbf{P}^{-1} (\mathbf{A} \mathbf{v}_1 \ \mathbf{A} \mathbf{v}_2 \ \dots \ \mathbf{A} \mathbf{v}_n) \quad (48.2.3)$$

$$= \mathbf{P}^{-1} (\lambda_1 \mathbf{v}_1 \ \lambda_2 \mathbf{v}_2 \ \dots \ \lambda_n \mathbf{v}_n) \quad (48.2.4)$$

$$= \mathbf{P}^{-1} \mathbf{P} \text{diag}(\lambda_i) \quad (48.2.5)$$

$$= \text{diag}(\lambda_i) \quad (48.2.6)$$

Thus, the new matrix in the basis \mathbf{P} is indeed diagonal, with entries equal to the eigenvalues. ■

Recall that the traces and determinants of a matrix are independent of the chosen basis, so if \mathbf{A} is diagonalizable then:

$$\text{tr} \mathbf{A} = \sum_i \lambda_i, \quad \det \mathbf{A} = \prod_i \lambda_i \quad (48.2.7)$$

This gives us a nice way to check if any arithmetic mistakes have been made in evaluating the eigenvalues.

We also gain some geometrical insight behind what eigenvectors really are. Indeed, if we consider any diagonalizable linear map f acting on vectors in V , it follows that its action

will be to stretch V by a factor of λ_i along \mathbf{v}_i . This can be readily verified by looking at the diagonalized form of f , and noting that the i th column of a matrix representation gives the vector that the corresponding i th basis vector gets mapped to.

In other words, the eigenvectors are vectors which, when acted upon by a linear map f , only change by a phase, but still point in the same "direction".

Example. Let's consider the one of the Pauli matrices

$$\sigma = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad (48.2.8)$$

Its characteristic equation is:

$$\chi_\sigma(\lambda) = \lambda^2 - 1 = 0 \implies \lambda = \pm 1 \quad (48.2.9)$$

For $\lambda_1 = 1$ we find that:

$$\begin{pmatrix} -1 & -i \\ i & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \implies x = -iy \quad (48.2.10)$$

Its eigenspace is:

$$\text{Eig}_\sigma(\lambda_1) = \left\{ k \begin{pmatrix} -i \\ 1 \end{pmatrix} : k \in \mathbb{R}^* \right\} \quad (48.2.11)$$

Similarly, for $\lambda_2 = -1$ we find that:

$$\begin{pmatrix} 1 & -i \\ i & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \implies x = iy \quad (48.2.12)$$

Its eigenspace is:

$$\text{Eig}_\sigma(\lambda_2) = \left\{ k \begin{pmatrix} i \\ 1 \end{pmatrix} : k \in \mathbb{R}^* \right\} \quad (48.2.13)$$

We therefore choose the eigenbasis with $k = 1$, whose transition matrix is:

$$P = \begin{pmatrix} -i & i \\ 1 & 1 \end{pmatrix} \quad (48.2.14)$$

whose inverse is (since $\det P = -i - i = -2i$):

$$P^{-1} = -\frac{1}{2i} \begin{pmatrix} 1 & -i \\ -1 & -i \end{pmatrix} = \frac{i}{2} \begin{pmatrix} 1 & -i \\ -1 & -i \end{pmatrix} = \frac{1}{2} \begin{pmatrix} i & 1 \\ -i & 1 \end{pmatrix} \quad (48.2.15)$$

We find that:

$$\sigma' = \frac{1}{2} \begin{pmatrix} i & 1 \\ -i & 1 \end{pmatrix} \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \begin{pmatrix} -i & i \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad (48.2.16)$$

◀

Theorem (Simultaneous diagonalization)

Two linear maps f, g are simultaneously diagonalizable, that is, they are diagonalized by the same matrix P , iff they commute:

$$[f, g] \equiv f \circ g - g \circ f = 0 \quad (48.2.17)$$

Proof. (\implies) Suppose that f, g are simultaneously diagonalizable, so that they share a set of eigenvectors \mathbf{v}_i with eigenvalues λ_i and λ'_i respectively. Then:

$$[f, g](\mathbf{v}_i) = \lambda_i g(\mathbf{v}_i) - \lambda'_i f(\mathbf{v}_i) = \lambda_i \lambda'_i - \lambda_i \lambda'_i = 0 \quad (48.2.18)$$

Any vector can be expanded as a linear combination of \mathbf{v}_i , so $[f, g](\mathbf{v}) = 0$ holds for all $\mathbf{v} \in V$. It follows that $[f, g] = 0$, the two maps commute.

(\impliedby) Suppose that f, g commute, and suppose that f has eigenvectors \mathbf{v}_i with eigenvalues λ_i . Then:

$$(f \circ g)(\mathbf{v}_i) = f(g(\mathbf{v}_i)) = (g \circ f)(\mathbf{v}_i) = \lambda_i g(\mathbf{v}_i) \quad (48.2.19)$$

implying that $g(\mathbf{v}_i) \in \text{Eig}_f(\lambda_i)$. For non-degenerate eigenvalues, this means that $g(\mathbf{v}_i) = \alpha \mathbf{v}_i$ for some non-zero α . Hence \mathbf{v}_i is an eigenvector of both f and g , the two maps are simultaneously diagonalizable. ■

48.3 Orthogonal diagonalization

Theorem (Spectral properties for hermitian matrices)

Let A be a hermitian matrix (so that $A^\dagger = A$). Then all its eigenvalues are real, and the eigenvectors corresponding to distinct eigenvalues are orthogonal.

Proof. Let \mathbf{v} be an eigenvector of A with eigenvalue λ . Then, it follows that

$$\langle \mathbf{v}, A\mathbf{v} \rangle = \lambda = \langle A\mathbf{v}, \mathbf{v} \rangle = \lambda^* \implies \lambda \in \mathbb{R} \quad (48.3.1)$$

With this established, an immediate consequence is that if $A\mathbf{v}_i = \lambda_i \mathbf{v}_i$ and $A\mathbf{v}_j = \lambda_j \mathbf{v}_j$ with $\lambda_i \neq \lambda_j$ then:

$$\langle \mathbf{v}_i, A\mathbf{v}_j \rangle = \lambda_j \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \lambda_i \langle \mathbf{v}_i, \mathbf{v}_j \rangle \quad (48.3.2)$$

giving:

$$(\lambda_j - \lambda_i) \langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0 \quad (48.3.3)$$

and since by assumption $\lambda_i \neq \lambda_j$ then this can only occur $\langle \mathbf{v}_i, \mathbf{v}_j \rangle$. ■

Interestingly, Hermitian matrices can always be diagonalized.

Proposition (Hermitian diagonalizability)

Let V be an n dimensional vector space over \mathbb{C} , if f is a Hermitian map defined on V then it has an orthonormal eigenbasis $\mathcal{E} = \{\mathbf{v}_i\}$.

Proof. We consider non-degenerate maps f , and proceed by induction.

If $\dim V = 1$, then the result is trivial.

Suppose for all $k < n$ we have shown that all hermitian maps have an orthonormal eigenbasis. Then, let's consider a map f on an n -dimensional V . We have that its characteristic equation must have at least one root λ over \mathbb{C} . Its eigenspace $W \equiv \text{Eig}_f(\lambda)$ is such that $\dim W > 0$, and we consider its orthogonal complement $W^\perp = \{\mathbf{u} : \in V : \langle \mathbf{u}, \mathbf{v}_i \rangle = 0 \forall \mathbf{v}_i \in W\}$. We have that for $\mathbf{u} \in W^\perp$:

$$\langle \mathbf{w}, f(\mathbf{u}) \rangle = \langle f(\mathbf{w}), \mathbf{u} \rangle = \lambda \langle \mathbf{w}, \mathbf{u} \rangle = 0 \quad (48.3.4)$$

so $f(\mathbf{u}) \in W^\perp$. Consequently, we may restrict f to W^\perp , and define its restriction as $g \equiv f|_{W^\perp}$. Since $\dim W^\perp = k < n$, we can use the induction hypothesis to deduce that it has an orthonormal eigenbasis $\{\mathbf{v}_i\}$. Furthermore, since $\dim W^\perp + \dim W = \dim V = n$, we have that $\{\mathbf{v}_i\} \cup W$ will give a set of n linearly independent eigenvectors of f , as desired. ■

Hermitian matrices play an important role, since their diagonalization is often easier to perform.

Theorem (Diagonalizing hermitian matrices) Let f be a hermitian linear map on V . Then, its diagonalized form is found through the similarity transformation:

$$\mathbf{A}' = \mathbf{P}^T \mathbf{A} \mathbf{P} \quad (48.3.5)$$

where \mathbf{A} is the matrix representation of f in some basis \mathcal{B} , and

$$\mathbf{P} = ([\mathbf{v}_1]_{\mathcal{B}} \ [\mathbf{v}_2]_{\mathcal{B}} \ \dots \ [\mathbf{v}_n]_{\mathcal{B}}) \quad (48.3.6)$$

Proof. We have that:

$$P^T P = \begin{pmatrix} [\mathbf{v}_1]_{\mathcal{B}}^T \\ [\mathbf{v}_2]_{\mathcal{B}}^T \\ \vdots \\ [\mathbf{v}_n]_{\mathcal{B}}^T \end{pmatrix} ([\mathbf{v}_1]_{\mathcal{B}} \; [\mathbf{v}_2]_{\mathcal{B}} \; \dots \; [\mathbf{v}_n]_{\mathcal{B}}) \quad (48.3.7)$$

$$= \begin{pmatrix} [\mathbf{v}_1]_{\mathcal{B}}^T [\mathbf{v}_1]_{\mathcal{B}} & [\mathbf{v}_1]_{\mathcal{B}}^T [\mathbf{v}_2]_{\mathcal{B}} & \dots & [\mathbf{v}_1]_{\mathcal{B}}^T [\mathbf{v}_n]_{\mathcal{B}} \\ [\mathbf{v}_2]_{\mathcal{B}}^T [\mathbf{v}_1]_{\mathcal{B}} & [\mathbf{v}_2]_{\mathcal{B}}^T [\mathbf{v}_2]_{\mathcal{B}} & \dots & [\mathbf{v}_2]_{\mathcal{B}}^T [\mathbf{v}_n]_{\mathcal{B}} \\ \vdots & \vdots & \ddots & \vdots \\ [\mathbf{v}_n]_{\mathcal{B}}^T [\mathbf{v}_1]_{\mathcal{B}} & [\mathbf{v}_n]_{\mathcal{B}}^T [\mathbf{v}_2]_{\mathcal{B}} & \dots & [\mathbf{v}_n]_{\mathcal{B}}^T [\mathbf{v}_n]_{\mathcal{B}} \end{pmatrix} \quad (48.3.8)$$

$$= \mathbb{I} \quad (48.3.9)$$

Consequently P is unitary, and hence when diagonalizing A' according to the general procedure:

$$A' = P^{-1}AP = P^TAP \quad (48.3.10)$$

as desired. ■

Example. Let's diagonalize the following hermitian matrix:

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \quad (48.3.11)$$

Its characteristic equation is:

$$\begin{vmatrix} 1 - \lambda & 0 & 0 \\ 0 & 2 - \lambda & 1 \\ 0 & 1 & 2 - \lambda \end{vmatrix} = (1 - \lambda)((2 - \lambda)^2 - 1) \quad (48.3.12)$$

$$= (1 - \lambda)(1 - \lambda)(3 - \lambda) = 0 \quad (48.3.13)$$

giving $\lambda = 1, 3$. It may seem like this matrix is not diagonalisable, since we only have two eigenvalues. However, we know that this can't be the case, Hermitian matrices are always diagonalisable. Indeed, although we only have two eigenvalues, it turns out that the first $\lambda_1 = 1$ will have a two dimensional eigenspace, so we will be able to find two orthonormal eigenvectors associated to this eigenvalue.

For $\lambda_1 = 1$ we get that

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \implies y = -z \quad (48.3.14)$$

This gives a two-dimensional eigenspace:

$$\text{Eig}_A(\lambda_1) = \left\{ t \begin{pmatrix} k \\ 1 \\ -1 \end{pmatrix}, \forall k, t \in \mathbb{R} \right\} \quad (48.3.15)$$

so we can choose as our orthonormal eigenvectors:

$$\mathbf{v}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad (48.3.16)$$

Finally, for $\lambda_2 = 3$ then we get that:

$$\begin{pmatrix} -2 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \implies x = 0, y = z \quad (48.3.17)$$

giving the following eigenspace:

$$\text{Eig}_A(\lambda_2) = \left\{ t \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \forall t \in \mathbb{R}^* \right\} \quad (48.3.18)$$

We choose the following eigenvector

$$\mathbf{v}_3 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \quad (48.3.19)$$

Hence, the orthonormal eigenbasis has a transition matrix:

$$P = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & \sqrt{2} & 0 \\ 1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix} \implies P^T = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 1 & -1 \\ \sqrt{2} & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix} \quad (48.3.20)$$

The diagonalized form of A is then:

$$\frac{1}{2} \begin{pmatrix} 0 & 1 & -1 \\ \sqrt{2} & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 0 & \sqrt{2} & 0 \\ 1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix} \quad (48.3.21)$$

as desired. This concludes our process of orthogonal diagonalization. ◀

Definition (Normal linear map)

Let $f : V \rightarrow V$ be a linear map. Then, f is normal iff $[f, f^\dagger] = 0$.

Note that unitary and hermitian maps are special cases of normal linear maps.

Proposition (Hermitian adjoint of normal map)

Let $f : V \rightarrow V$ be a normal linear map over V , and let $\mathbf{v} \in \text{Eig}_f(\lambda) \subseteq V$ be an eigenvector of f with eigenvalue λ . Then, we have that $f^\dagger(\mathbf{v}) = \lambda^* \mathbf{v}$.

Proof. Let $g = f - \lambda \text{id}$. Then we find that:

$$g \circ g^\dagger = (f - \lambda \text{id}) \circ (f^\dagger - \lambda^* \text{id}) = f \circ f^\dagger - \lambda^* f - \lambda f^\dagger + |\lambda|^2 \text{id} \quad (48.3.22)$$

$$= f^\dagger \circ f - \lambda^* f - \lambda f^\dagger + |\lambda|^2 \text{id} \quad (48.3.23)$$

$$= g^\dagger \circ g \quad (48.3.24)$$

implying that g is a normal linear map. Consequently:

$$0 = \langle g(\mathbf{v}), g(\mathbf{v}) \rangle = \langle \mathbf{v}, g^\dagger \circ g(\mathbf{v}) \rangle = \langle \mathbf{v}, g \circ g^\dagger(\mathbf{v}) \rangle = \langle g^\dagger(\mathbf{v}), g^\dagger(\mathbf{v}) \rangle \quad (48.3.25)$$

implying that $g^\dagger(\mathbf{v}) = f^\dagger(\mathbf{v}) - \lambda^* \mathbf{v} = 0$ as desired. ■

Theorem (Spectral theorem for normal maps)

Let $f : V \rightarrow V$ be a linear map. Then f is normal iff it has an orthonormal eigenvector basis, that is, it is diagonalizable.

Proof. (\implies) Suppose f is a normal linear map on V , we proceed by induction.

If $\dim V = n = 1$, then the result is trivially verified.

Suppose that for $\dim V = k < n$ we have shown that all normal linear maps have an orthonormal eigenvector basis. Then, let's consider a normal linear map f on an n -dimensional V . We have that its characteristic equation must have at least one root λ over \mathbb{C} . Its eigenspace $W \equiv \text{Eig}_f(\lambda)$ is such that $\dim W = 1$, and we consider its orthogonal complement $W^\perp = \{\mathbf{u} : \langle \mathbf{u} \in V : \mathbf{u}, \mathbf{v}_i \rangle = 0 \forall \mathbf{v}_i \in W\}$ with $\dim W^\perp = n - 1$. We have that for $\mathbf{u} \in W^\perp$:

$$\langle f(\mathbf{u}), \mathbf{v} \rangle = \langle \mathbf{u}, f^\dagger(\mathbf{v}) \rangle = \lambda^* \langle \mathbf{w}, \mathbf{v} \rangle = 0 \quad (48.3.26)$$

$$\langle f^\dagger(\mathbf{u}), \mathbf{v} \rangle = \langle \mathbf{u}, f(\mathbf{v}) \rangle = \lambda \langle \mathbf{w}, \mathbf{v} \rangle = 0 \quad (48.3.27)$$

These two results imply that $f(W^\perp), f^\dagger(W^\perp) \subseteq W$, and we may therefore consider the restriction $f|_{W^\perp}$. By the induction assumption, this normal map has $n - 1$ orthonor-

mal eigenvectors. Adding $\frac{\mathbf{v}}{|\mathbf{v}|}$ gives the desired list of n linearly independent orthonormal eigenvectors.

(\Leftarrow) Suppose that f has an orthonormal eigenvector basis $\{\mathbf{n}_i\}$. Then:

$$f \circ f^\dagger(\mathbf{n}_i) = \lambda_i^* f(\mathbf{n}_i) = |\lambda_i|^2 \mathbf{n}_i \quad (48.3.28)$$

and

$$f^\dagger \circ f(\mathbf{n}_i) = \lambda_i f(\mathbf{n}_i) = |\lambda_i|^2 \mathbf{n}_i \quad (48.3.29)$$

implying that

$$[f, f^\dagger](\mathbf{n}_i) = 0 \quad (48.3.30)$$

Given any vector $\mathbf{v} \in V$, it may be expanded in the eigenbasis as $\mathbf{v} = \sum_i \alpha_i \mathbf{n}_i$ so that:

$$[f, f^\dagger](\mathbf{v}) = \sum_i \alpha_i [f, f^\dagger](\mathbf{n}_i) = 0, \forall \mathbf{v} \in V \implies [f, f^\dagger] = 0 \quad (48.3.31)$$

as desired. ■

Notice the resemblance between this proof and the proof that all hermitian maps are diagonalizable. We proceeded by showing that there must be some eigenvector, and that its orthogonal complement is invariant under the map we are interested in. Diagonalizing the restriction of the map to the orthogonal complement gives an extra set of eigenvectors which we can use to complete the proof.

48.4 Classifying conics

Suppose we have a conic with general equation:

$$Ax^2 + Bxy + Cy^2 + Fx + Gy + H = 0 \quad (48.4.1)$$

Our goal will be to classify this conic as either a parabola, hyperbola or ellipse, and determine some of its fundamental features.

Aligning the axes

We can write (48.4.1) as a product of matrices:

$$\begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} A & \frac{B}{2} \\ \frac{B}{2} & C \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} F & G \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + H = 0 \quad (48.4.2)$$

Let us define:

$$\mathbf{A} = \begin{pmatrix} A & \frac{B}{2} \\ \frac{B}{2} & C \end{pmatrix}, \mathbf{J} = \begin{pmatrix} F & G \end{pmatrix}, \mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix} \quad (48.4.3)$$

then (48.4.2) turns into

$$\mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{J}^T \mathbf{x} + H = 0 \quad (48.4.4)$$

It is important to note that A is a symmetric matrix, and can therefore be orthogonally diagonalized. Suppose that $D = P^T A P$, then we get that:

$$\mathbf{x}^T P D P^T \mathbf{x} + \mathbf{J}^T \mathbf{x} + H = 0 \quad (48.4.5)$$

Let us define the coordinate vectors in the eigenbasis as: $\mathbf{x}' = P^T \mathbf{x}$ (recall that P represents a change from the eigenbasis, so its inverse/transpose will represent a change to the eigenbasis). Then we find that:

$$(\mathbf{x}')^T D \mathbf{x}' + \mathbf{J}^T P \mathbf{x}' + H = 0 \quad (48.4.6)$$

The process we have gone through can be viewed geometrically as rotating \mathbb{R}^2 to align the axes with the eigenvector basis of A . Indeed, since A is symmetric, its transition matrix will be unitary, it will represent a rotation/reflection. By performing a change of basis $\mathbf{x} \rightarrow \mathbf{x}'$ we were really just rotating \mathbb{R}^2 . Suppose $D = \text{diag}(\lambda_1, \lambda_2)$ and $\mathbf{J}^T P = (f \ g)$, then we find:

$$\lambda_1 x'^2 + \lambda_2 y'^2 + f x' + g y' + H = 0 \quad (48.4.7)$$

Translating the origin

The final step is translating the origin to get a standard conic. For ellipses and hyperbolas we do so by completing the square:

$$\lambda_1 x'^2 + \lambda_2 y'^2 + f x' + g y' + H = 0 \quad (48.4.8)$$

$$\implies \lambda_1 \left(x' + \frac{f}{2\lambda_1} \right)^2 + \lambda_2 \left(y' + \frac{g}{2\lambda_2} \right)^2 + H - \frac{f^2}{4\lambda_1} - \frac{g^2}{4\lambda_2} = 0 \quad (48.4.9)$$

Letting the translated axes be defined by:

$$\mathbf{x}'' = \mathbf{x}' + \frac{1}{2} \begin{pmatrix} \frac{f}{\lambda_1} \\ \frac{g}{\lambda_2} \end{pmatrix} \quad (48.4.10)$$

then we find that:

$$\lambda_1 x''^2 + \lambda_2 y''^2 + h = 0, \quad h = H - \frac{f^2}{4\lambda_1} - \frac{g^2}{4\lambda_2} \quad (48.4.11)$$

which is a conic. It can be rearranged into the more useful form :

$$\frac{x''^2}{a^2} + \frac{y''^2}{b^2} = 1 \quad (48.4.12)$$

where $a^2 = -\frac{h}{\lambda_1}$ and $b^2 = -\frac{h}{\lambda_2}$. Depending on the values of a, b this will be either a hyperbola or ellipse.

If instead we are dealing with a parabola, then we will find that one of the eigenvalues is

Conic	Standard form
Hyperbola	$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$
Parabola	$y^2 - 4ax = 0$
Ellipse	$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$

0. Suppose WLOG that $\lambda_1 = 0$, then we find:

$$\lambda_2 y'^2 + f x' + g y' + H = 0 \quad (48.4.13)$$

$$\implies \lambda_2 \left(y' + \frac{g}{2\lambda_2} \right)^2 + f x' + H - \frac{g^2}{4\lambda_2} = 0 \quad (48.4.14)$$

Letting the translated axes be defined by:

$$\mathbf{x}'' = \mathbf{x}' + \begin{pmatrix} \frac{H}{f} - \frac{g^2}{4f\lambda_2} \\ \frac{g}{2\lambda_2} \end{pmatrix} \quad (48.4.15)$$

then we find that:

$$y''^2 + \frac{f}{\lambda_2} x'' = 0 \quad (48.4.16)$$

which is a parabola. Had $\lambda_2 = 0$ then we would have found in complete analogy to before:

$$x''^2 + \frac{g}{\lambda_1} y'' = 0 \quad (48.4.17)$$

Example. Let's classify the conic described by:

$$x^2 - 4xy + 4y^2 - 6x - 8y + 5 = 0 \quad (48.4.18)$$

which may be re-expressed in matrix form as:

$$\mathbf{x}^T \begin{pmatrix} 1 & -2 \\ -2 & 4 \end{pmatrix} \mathbf{x} + (-6 \ -8) \mathbf{x} + 5 = 0 \quad (48.4.19)$$

The eigenvalues of \mathbf{A} are easily found to obey

$$(1 - \lambda)(4 - \lambda) - 4 = 0 \implies \lambda(\lambda - 5) = 0 \implies \lambda_1 = 0, \lambda_2 = 5 \quad (48.4.20)$$

For $\lambda_1 = 0$ we get the eigenspace

$$\text{Eig}_{\mathbf{A}}(\lambda_1) = \left\{ k \begin{pmatrix} 2 \\ 1 \end{pmatrix} : k \in \mathbb{R}^* \right\} \quad (48.4.21)$$

Similarly, for $\lambda_2 = 5$ we get the eigenspace:

$$\text{Eig}_A(\lambda_2) = \left\{ k \begin{pmatrix} 1 \\ -2 \end{pmatrix} : k \in \mathbb{R}^* \right\} \quad (48.4.22)$$

so we may choose the orthonormal basis $\left\{ \frac{1}{\sqrt{5}} \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ -2 \end{pmatrix} \right\}$ with transition matrix:

$$P = \frac{1}{\sqrt{5}} \begin{pmatrix} 2 & 1 \\ 1 & -2 \end{pmatrix} \implies J^T P = \frac{1}{\sqrt{5}} \begin{pmatrix} -20 & 10 \end{pmatrix} \quad (48.4.23)$$

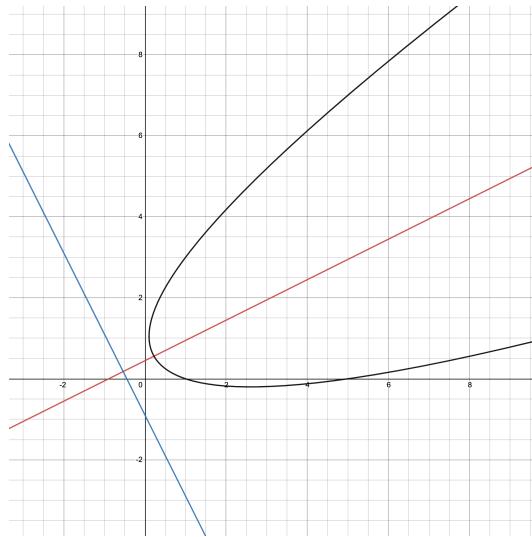
Consequently we find:

$$5y'^2 - 4\sqrt{5}x' + 2\sqrt{5}y' + 5 = 0 \implies y'^2 - \frac{4\sqrt{5}}{5}x' + \frac{2\sqrt{5}}{5}y' + 1 = 0 \quad (48.4.24)$$

We now complete the square:

$$(y' + \frac{\sqrt{5}}{5})^2 - \frac{4\sqrt{5}}{5}x' + \frac{4}{5} = 0 \implies y''^2 = \frac{4\sqrt{5}}{5}x'' \quad (48.4.25)$$

where $y'' = y' + \frac{\sqrt{5}}{5}$ and $x'' = x' - \frac{\sqrt{5}}{5}$. This is therefore a parabola.



48.5 Matrix exponentials and Lie algebras

One final important application of diagonalization is in determining matrix exponents.

Proposition (Matrix exponents)

Suppose that A is a diagonalizable matrix with $A' = P^{-1}AP$. Then:

$$A^n = PA'^n P^{-1} \quad (48.5.1)$$

Proof. This follows immediately from:

$$A^n = \underbrace{(PAP^{-1})(PAP^{-1})\dots(PAP^{-1})}_{n \text{ times}} = PA'^n P^{-1} \quad (48.5.2)$$

■

Definition (Matrix exponential)

Suppose that A matrix, then its exponential is defined as:

$$e^A = \sum_{i=0}^{\infty} \frac{1}{n!} A^n = \mathbb{1} + A + \frac{1}{2}A^2 + \dots \quad (48.5.3)$$

48.6 Schur's triangulation theorem

Theorem (Schur's triangulation theorem)

Let $A \in \text{Mat}_n(\mathbb{C})$ with eigenvalues $\lambda_1, \lambda_2, \dots$ which may be degenerate.. Then A is unitarily equivalent to an upper triangular matrix:

$$A = UTU^\dagger \quad (48.6.1)$$

where:

$$T = \begin{pmatrix} \lambda_1 & & \dots \\ 0 & \lambda_2 & & \\ \vdots & & \ddots & \end{pmatrix} \quad (48.6.2)$$

Proof. We proceed by induction. For $n = 1$ the result is trivial. Suppose we have shown that any $m \times m$ matrix where $m \leq n - 1$ is unitarily equivalent to an upper triangular matrix. Let $A \in \text{Mat}_n(\mathbb{C})$ have eigenvalues $\lambda_1, \lambda_2, \dots$ and eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \dots$, which may be degenerate, and are assumed to have unit norm.

Now \mathbf{v}_1 can be used to form an orthonormal basis $\{\mathbf{v}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$. Then, the resulting

matrix when we change to this basis will be unitarily equivalent to \mathbf{A} :

$$\mathbf{A} = \mathbf{V} \begin{pmatrix} \lambda_1 & a_{12} & \dots & a_{1n} \\ 0 & & & \\ \vdots & & \tilde{\mathbf{A}} & \\ 0 & & & \end{pmatrix} \mathbf{V}^\dagger \quad (48.6.3)$$

Clearly, we must have that $\chi_{\mathbf{A}}(\lambda) = (\lambda_1 - \lambda)\chi_{\tilde{\mathbf{A}}}(\lambda)$, implying that $\tilde{\mathbf{A}}$ has eigenvalues $\lambda_2, \lambda_3, \dots$ identical to \mathbf{A} , and which could be degenerate. We can now use the induction hypothesis, since $\tilde{\mathbf{A}} \in \text{Mat}_{n-1}(\mathbb{C})$ we have that it is unitarily equivalent to some upper triangular matrix:

$$\tilde{\mathbf{A}} = \tilde{\mathbf{W}} \begin{pmatrix} \lambda_2 & \tilde{a}_{12} & \dots & \tilde{a}_{1n} \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \tilde{a}_{n-1n} \\ 0 & \dots & 0 & \lambda_n \end{pmatrix} \tilde{\mathbf{W}}^\dagger \quad (48.6.4)$$

Note also that:

$$\left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \tilde{\mathbf{W}} & \\ 0 & & & \end{array} \right)^\dagger \left(\begin{array}{c|ccc} \lambda_1 & a_{12} & \dots & a_{1n} \\ \hline 0 & & & \\ \vdots & & \tilde{\mathbf{A}} & \\ 0 & & & \end{array} \right) \left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \tilde{\mathbf{W}} & \\ 0 & & & \end{array} \right) = \left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \tilde{\mathbf{W}}^\dagger & \\ 0 & & & \end{array} \right) \left(\begin{array}{c|ccc} \lambda_1 & b_{12} & \dots & b_{1n} \\ \hline 0 & & & \\ \vdots & & \tilde{\mathbf{A}} \tilde{\mathbf{W}} & \\ 0 & & & \end{array} \right) \quad (48.6.5)$$

$$= \left(\begin{array}{c|ccc} \lambda_1 & c_{12} & \dots & c_{1n} \\ \hline 0 & & & \\ \vdots & & \tilde{\mathbf{W}}^\dagger \tilde{\mathbf{A}} \tilde{\mathbf{W}} & \\ 0 & & & \end{array} \right) = \left(\begin{array}{ccccc} \lambda_1 & \tilde{b}_{12} & \dots & \dots & \tilde{b}_{1n} \\ 0 & \lambda_2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \tilde{a}_{n-1n} \\ 0 & \dots & \dots & 0 & \lambda_n \end{array} \right) \quad (48.6.6)$$

Defining:

$$\mathbf{W} = \left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \tilde{\mathbf{W}} & \\ 0 & & & \end{array} \right) \quad (48.6.7)$$

we see that \mathbf{W} is unitary and that

$$\mathbf{A} = \mathbf{V} \mathbf{W} \begin{pmatrix} \lambda_1 & a_{12} & \dots & a_{1n} \\ \hline 0 & & & \\ \vdots & & \tilde{\mathbf{A}} & \\ 0 & & & \end{pmatrix} \mathbf{W}^\dagger \mathbf{V}^\dagger \quad (48.6.8)$$

Thus A is unitarily triangularizable via $U = VW$. ■

Theorem (The Cayley-Hamilton theorem)

Let $A \in \text{Mat}_n(\mathbb{C})$ have characteristic polynomial $\chi_A(\lambda)$. Then, we have that $\chi_A(A) = 0$ where 0 is the zero element of $\text{Mat}_n(\mathbb{C})$,

Proof. We can factorize the characteristic polynomial into the following form due to the Fundamental theorem of algebra:

$$\chi_A(\lambda) = (\lambda_1 - \lambda)(\lambda_2 - \lambda)\dots(\lambda_n - \lambda) \quad (48.6.9)$$

implying that

$$\chi_A(A) = (\lambda_1 \mathbb{1} - A)(\lambda_2 \mathbb{1} - A)\dots(\lambda_n \mathbb{1} - A) \quad (48.6.10)$$

Since $A \in \text{Mat}_n(\mathbb{C})$, Schur's theorem tells us that it can be triangularized unitarily $A = UTU^\dagger$ where T is upper triangular. Therefore:

$$\chi_A(A) = U(\lambda_1 \mathbb{1} - T)U^\dagger U(\lambda_2 \mathbb{1} - T)U^\dagger \dots U(\lambda_n \mathbb{1} - T)U^\dagger \quad (48.6.11)$$

$$= U(\lambda_1 \mathbb{1} - T)(\lambda_2 \mathbb{1} - T)\dots(\lambda_n \mathbb{1} - T)U^\dagger \quad (48.6.12)$$

Each of the factors $\lambda_i \mathbb{1} - T$ will be upper triangular with the i th diagonal element equal to zero. It is easy to verify that a product of such matrices must be null. Let $A = \lambda_1 \mathbb{1} - T$ and $B = \lambda_2 \mathbb{1} - T$, and define $C = AB$. We have that $C_{11} = C_{22} = 0$ since in general for triangular matrices:

$$C_{ii} = \sum_j A_{ij}B_{ji} = A_{ii}B_{ii} \quad (48.6.13)$$

Instead, $C_{12} = \sum_j A_{1j}B_{j2} = 0$ since A_{11} and B_{22} are both zero. The first two rows of C are thus equal to zero.

Suppose we have repeated this process up to the factor $(\lambda_m \mathbb{1} - T)$ so that the first $m-1$ rows are all zero. Then letting $C = \prod_{i=1}^m (\lambda_i \mathbb{1} - T) = A(\lambda_m \mathbb{1} - T)$ we get:

$$C_{lm} = \sum_{l \leq j \leq m} A_{lj}B_{jm} \quad (48.6.14)$$

Now since $A_{lj} = 0$ for all $j < m$, the only term that will survive will be that with $j = m$. Consequently:

$$C_{lm} = A_{lm}B_{mm} = 0 \quad (48.6.15)$$

so the m th column will also be zero. It follows by induction that $(\lambda_1 \mathbb{1} - A)\dots(\lambda_n \mathbb{1} - A) = 0$ and thus $\chi_A(A) = 0$ as desired. ■

We can use Schur's triangulation theorem to prove the Spectral theorem more generally.

Proposition (Normal triangular matrices) A triangular matrix is normal iff it is diagonal.

Proof. The \Leftarrow is trivial. Suppose $A \in \text{Mat}_n(\mathbb{C})$ is a normal triangular matrix. The case $n = 1$ is obvious. Suppose the proposition is true for all $m \times m$ matrices where $m \leq n - 1$. Then, writing A as:

$$A = \begin{pmatrix} a_{11} & \mathbf{a} \\ 0 & \tilde{A} \end{pmatrix} \implies A^\dagger = \begin{pmatrix} a_{11}^* & 0 \\ \mathbf{a}^* & \tilde{A}^\dagger \end{pmatrix} \quad (48.6.16)$$

we see that

$$A^\dagger A = \begin{pmatrix} |a_{11}|^2 & \dots \\ \dots & \tilde{A}^\dagger \tilde{A} + \|\mathbf{a}\|^2 \end{pmatrix} \quad (48.6.17)$$

$$AA^\dagger = \begin{pmatrix} |a_{11}|^2 + \|\mathbf{a}\|^2 & \dots \\ \dots & \tilde{A} \tilde{A}^\dagger \end{pmatrix} \quad (48.6.18)$$

For these to be equal, we need $a_{11} = 0$, $\mathbf{a} = 0$ and $\tilde{A}^\dagger \tilde{A} = \tilde{A} \tilde{A}^\dagger$. Also, since $A \in \text{Mat}_{n-1}(\mathbb{C})$ is upper triangular and normal, it must be diagonal. Consequently A is diagonal, as required. ■

Theorem (The Spectral Theorem) Let $A \in \text{Mat}_n(\mathbb{C})$ with eigenvalues $\lambda_1, \dots, \lambda_n$, the following are equivalent:

- (i) A is normal
- (ii) A is unitarily diagonalizable
- (iii) $\sum_{ij} |A_{ij}|^2 = \sum_i |\lambda_i|^2$

Proof.

(i) \implies (ii) If A is unitarily diagonalizable then $A = UDU^\dagger$ and thus

$$A^\dagger A = (UD^\dagger U^\dagger)(UDU^\dagger) = U D^\dagger D U^\dagger \quad (48.6.19)$$

$$AA^\dagger = (UDU^\dagger)(UD^\dagger U^\dagger) = U D D^\dagger U^\dagger \quad (48.6.20)$$

but $D^\dagger D = DD^\dagger$ so A is normal.

(ii) \implies (i) Suppose A is normal. By Schur's theorem it is unitarily equivalent to an upper triangular matrix.

Lemma. Normality is preserved under unitary transformations.

Indeed if $A^\dagger A = AA^\dagger$ then

$$(UAU^\dagger)^\dagger(UAU^\dagger) = UA^\dagger AU^\dagger = UAA^\dagger U^\dagger = (UAU^\dagger)(UAU^\dagger)^\dagger \quad (48.6.21)$$

Consequently, the upper triangular decomposition of A must be normal, and thus diagonal, as desired.

(ii) \implies (iii) Since A is unitarily diagonalizable, we have that $A = UDU^\dagger$ where $D = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ and thus:

$$\sum_{ij} |a_{ij}|^2 = \text{tr}(A^\dagger A) = \text{tr}(UD^\dagger DU^\dagger) = \text{tr}(U^\dagger U) \text{tr}(D) = \sum_i |\lambda_i|^2 \quad (48.6.22)$$

(iii) \implies (ii) By Schur's theorem we have that A is triangularizable into T with $\sum_i |T_{ii}|^2 = \sum_i |\lambda_i|^2$. Note also that (iii) implies:

$$\sum_{ij} |a_{ij}|^2 = \text{tr}(A^\dagger A) = \text{tr}(T^\dagger T) = \sum_{ij} |T_{ij}|^2 = \sum_i |\lambda_i|^2 \quad (48.6.23)$$

We therefore find that $\sum_i |T_{ii}|^2 = \sum_{ij} |T_{ij}|^2$ which is only possible if T is diagonal. ■

48.7 Jordan canonical form (make sure to write by October)

Tensor algebra

We have studied several types of mathematical objects until now (especially in linear algebra), and have seen that many tend to transform under changes of basis in different ways. Scalars are invariant under basis changes, whereas vectors transform under matrix multiplication. Linear operators instead transform using similarity transformations.

Tensor algebra studies the way we may categorize the ways objects transform under changes of basis, and the properties that follow from such classifications. In the next chapter on Tensor calculus, we study how we can differentiate these objects.

49.1 Einstein summation convention

Consider a vector space V with a basis $\mathcal{B} = \{v_1, v_2, \dots, v_n\}$, which we modify to $\mathcal{B}' = \{v'_1, v'_2, \dots, v'_n\}$. From our study of linear algebra we know that we can express this change of basis as:

$$\begin{pmatrix} v'_1 \\ \vdots \\ v'_n \end{pmatrix} = \begin{pmatrix} A_{11} & \dots & A_{1n} \\ \dots & \dots & \dots \\ A_{n1} & \dots & A_{nn} \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \quad (49.1.1)$$

Alternatively:

$$v'_i = \sum_{j=1}^n A_{ij} v_j \quad (49.1.2)$$

for $1 \leq i \leq n$. Here, we call the index i as the **running index**, whereas the index j is called the **dummy index**.

Often times calculations in fields such as General relativity or QFT require the use of several dummy indices, and could lead to a clutter of summation symbols Σ . Indeed, Einstein himself faced this problem when trying to work out the differential geometry of his theory of space-time, and to fix this issue, he devised a notation known as **Einstein notation**.

In this notation, if an index appears more than once in a summation, then the corresponding

\sum symbol may be omitted, since summation is implied. In other words, we would find that:

$$v'_i = \sum_{j=1}^n A_{ij} v_j \longrightarrow A_{ij} v_j \quad (49.1.3)$$

Suppose we wish to calculate the product of three matrices:

$$(ABC)_{il} = \sum_{j=1}^n \sum_{k=1}^n A_{ij} B_{jk} C_{kl} \quad (49.1.4)$$

In einstein notation, this becomes:

$$(ABC)_{il} = A_{ij} B_{jk} C_{kl} \quad (49.1.5)$$

which is considerably shorter.

In general, the greek alphabet is reserved for indices of space time components only. These indices therefore only take values of 0, 1, 2, 3..., where 0 is the temporal component. Instead, the normal alphabet is reserved for indices of spatial components only, so these indices run from 1, 2, 3.... according to the dimension of the space we are working in.

49.2 Cartesian tensors

A cartesian coordinate system in an n-dimensional space associates a coordinate (x_1, x_2, \dots, x_n) to each point in this space, with reference to a set of n basis vectors.

Suppose we have some vector \mathbf{r} in this space, whose components are (x_1, x_2, x_3) in the cartesian system \mathcal{C} and (x'_1, x'_2, x'_3) in the primed cartesian system \mathcal{C}' . The former has a basis $\{\mathbf{e}_i\}$ and the latter has a basis $\{\mathbf{e}'_i\}$.

Definition (*Rigid rotation*)

A **rigid rotation** of the cartesian axes is the transformation of components of one cartesian system to another:

$$x'_j = R_{jk} x_k \quad (49.2.1a)$$

$$x_k = R_{jk} x'_j \quad (49.2.1b)$$

where $R_{jk} = \mathbf{e}'_j \cdot \mathbf{e}_k$.

Indeed, a general vector \mathbf{r} may be expressed as:

$$\mathbf{r} = x_k \mathbf{e}_k = x'_j \mathbf{e}'_j \quad (49.2.2)$$

and since $\mathbf{e}_k = (\mathbf{e}'_j \cdot \mathbf{e}_k)\mathbf{e}'_j = R_{jk}\mathbf{e}'_j$ we find that:

$$x_k R_{jk} \mathbf{e}'_j = x'_j \mathbf{e}'_j \implies x_k R_{jk} = x'_j \quad (49.2.3)$$

since $\{\mathbf{e}'_j\}$ is linearly independent. Similarly, we may write $\mathbf{e}'_j = (\mathbf{e}'_j \cdot \mathbf{e}_k)\mathbf{e}_k = R_{jk}\mathbf{e}_k$ so that:

$$x'_j \mathbf{e}'_j = x'_j R_{jk} \mathbf{e}_k = x_k \mathbf{e}_k \implies x_k R_{jk} = x'_j \quad (49.2.4)$$

as desired.

Example. Suppose we want to rotate the cartesian system with basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ by some angle ϕ about the \mathbf{e}_3 (anti-clockwise when viewed from the tip of \mathbf{e}_3), to get another cartesian system with basis $\{\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3\}$ then:

$$x'_1 = (\mathbf{e}_1 \cdot \mathbf{e}'_1)x_1 + (\mathbf{e}_2 \cdot \mathbf{e}'_1)x_2 + (\mathbf{e}_3 \cdot \mathbf{e}'_1)x_3 = \cos \phi x_1 + \sin \theta x_2 \quad (49.2.5)$$

$$x'_2 = (\mathbf{e}_1 \cdot \mathbf{e}'_2)x_1 + (\mathbf{e}_2 \cdot \mathbf{e}'_2)x_2 + (\mathbf{e}_3 \cdot \mathbf{e}'_2)x_3 = -\sin \phi x_1 + \cos \theta x_2 \quad (49.2.6)$$

$$x'_3 = (\mathbf{e}_1 \cdot \mathbf{e}'_3)x_1 + (\mathbf{e}_2 \cdot \mathbf{e}'_3)x_2 + (\mathbf{e}_3 \cdot \mathbf{e}'_3)x_3 = x_3 \quad (49.2.7)$$

Therefore, we define the **rotation vector**:

$$\mathbf{R} = \begin{pmatrix} \cos \phi & \sin \phi & 0 \\ -\sin \phi & \cos \phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (49.2.8)$$

so that a vector \mathbf{x} :

$$\mathbf{x}' = \mathbf{R}\mathbf{x} \quad (49.2.9)$$

◀

Theorem (Rotation operator \mathbf{R})

The rotation operator in three dimensions, defined as:

$$\mathbf{R} = \begin{pmatrix} \cos \phi & \sin \phi & 0 \\ -\sin \phi & \cos \phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (49.2.10)$$

is unitary/orthogonal, satisfying $\mathbf{R}^T \mathbf{R} = \mathbf{R} \mathbf{R}^T = \mathbb{I}$ and $\det \mathbf{R} = \pm 1$.

Proof. We can write that:

$$\mathbf{x}' = \mathbf{R}\mathbf{x} \iff \mathbf{x} = \mathbf{R}^{-1}\mathbf{x}' \quad (49.2.11)$$

or in component form:

$$x'_i = R_{ij}x_j \iff x_j = (\mathbf{R}^{-1})_{ji}x_i \quad (49.2.12)$$

implying that $(\mathbf{R}^{-1})_{ji} = R_{ij}$, or in other words:

$$\boxed{\mathbf{R}^T = \mathbf{R}^{-1}} \quad (49.2.13)$$

In other words, the rotation matrix \mathbf{R} is unitary, and in hindsight it is quite obvious why it should be unitary. A rigid cartesian rotation, as the name suggests, is rigid, and hence will preserve angles between two vectors (we will prove this soon).

Immediately, we also find that:

$$1 = \det \mathbb{I}_n = \det(\mathbf{R}\mathbf{R}^T) = \det \mathbf{R} \det \mathbf{R}^T = (\det \mathbf{R})^2 \quad (49.2.14)$$

so that:

$$\boxed{\det \mathbf{R} = \pm 1} \quad (49.2.15)$$

Another important property of rigid rotation matrices is that their components are unitary:

$$\boxed{R_{ik}R_{jk} = \delta_{ij}, \quad R_{ik}R_{il} = \delta_{kl}} \quad (49.2.16)$$

Indeed:

$$\mathbf{e}'_i \cdot \mathbf{e}'_j = (\mathbf{e}'_i \cdot \mathbf{e}_k)(\mathbf{e}'_j \cdot \mathbf{e}_k)\mathbf{e}_k \cdot \mathbf{e}_k = R_{ik}R_{jk} = \delta_{ij} \quad (49.2.17)$$

Similarly:

$$\mathbf{e}_k \cdot \mathbf{e}_l = (\mathbf{e}'_i \cdot \mathbf{e}_k)(\mathbf{e}'_i \cdot \mathbf{e}_l)\mathbf{e}'_i \cdot \mathbf{e}'_i = R_{ik}R_{il} = \delta_{kl} \quad (49.2.18)$$

as desired.

We could also see this by using the unitarity of \mathbf{R} and write:

$$\mathbb{I} = \mathbf{R}\mathbf{R}^{-1} = \mathbf{R}\mathbf{R}^T \implies \delta_{ij} = R_{ik}R_{jk} \quad (49.2.19)$$

■

Geometrically, we see that this result makes sense. Indeed R_{ik} and R_{jk} are the angles between $\mathbf{e}'_i, \mathbf{e}_k$ and $\mathbf{e}'_j, \mathbf{e}_k$. Since \mathbf{e}'_i and \mathbf{e}'_j are orthogonal, we cannot have that both terms in the product $R_{ik}R_{jk}$ be non-zero unless $i = j$.

Definition (*First order Cartesian tensor*)

A **first order Cartesian tensor** (or vector) is defined as a geometric object \mathbf{v} represented by the components v_i in the cartesian system \mathcal{C} and represented by the components v'_i in the cartesian system \mathcal{C}' , such that they transform under rigid cartesian rotations as :

$$\boxed{v'_i = R_{ij}v_j} \quad (49.2.20a)$$

$$\boxed{v_i = R_{ki}v'_k} \quad (49.2.20b)$$

where \mathbf{R} is the rotation matrix as defined (49.2.8).

Example. Consider the quantity $\mathbf{v} = (x_1^2, x_2^2)$, which transforms under rotations as:

$$v'_1 = (x'_1)^2 = (x_1 \cos \theta + x_2 \sin \theta)^2 \quad (49.2.21)$$

$$v'_2 = (x'_2)^2 = (-x_1 \sin \theta + x_2 \cos \theta)^2 \quad (49.2.22)$$

If this were a first order cartesian tensor, we would need:

$$v'_1 = v_1 \cos \theta + v_2 \sin \theta = x_1^2 \cos \theta + x_2^2 \sin \theta \quad (49.2.23)$$

$$v'_2 = -v_1 \sin \theta + v_2 \cos \theta = -x_1^2 \sin \theta + x_2^2 \cos \theta \quad (49.2.24)$$

which clearly isn't the case. Consequently this is not a first order cartesian tensor. Consider instead the quantity $\mathbf{u} = (x_2, -x_1)$, which transforms under rotations as:

$$u'_1 = x'_2 = -x_1 \sin \theta + x_2 \cos \theta \quad (49.2.25)$$

$$u'_2 = -x'_1 = x_1 \cos \theta + x_2 \sin \theta \quad (49.2.26)$$

If this were a first order cartesian tensor, we would need:

$$u'_1 = u_1 \cos \theta + u_2 \sin \theta = x_2 \cos \theta - x_1 \sin \theta \quad (49.2.27)$$

$$u'_2 = -u_1 \sin \theta + u_2 \cos \theta = x_2 \sin \theta - x_1 \cos \theta \quad (49.2.28)$$

which is true for all θ . Consequently \mathbf{u} is a first order cartesian tensor. ◀

Proposition (Scalar product invariance)

The scalar product of two vectors, $\mathbf{u} \cdot \mathbf{v}$, is invariant under rotations.

Proof. We consider:

$$u'_i v'_i = R_{ij} u_j R_{ik} u_k = R_{ij} R_{ik} u_j v_k = \delta_{jk} u_j v_k = u_j v_j \quad (49.2.29)$$

as desired. Hence the scalar product of first order cartesian tensors is a zeroth order cartesian tensor. ■

The definition of a second order cartesian tensor is quite similar to that of a first order cartesian tensor, only that rotations must be repeated twice due to the presence of two indices.

Definition (Second order Cartesian tensor)

A **second order Cartesian tensor** is defined as a geometric object T represented by the components T_{ij} in the cartesian system \mathcal{C} and represented by the components T'_{ij} in the cartesian system \mathcal{C}' , such that they transform under rigid cartesian rotations:

$$T'_{ij} = R_{ik}R_{jl}T_{kl} \quad (49.2.30a)$$

$$T_{kl} = R_{mk}R_{nl}T_{mn} \quad (49.2.30b)$$

or alternatively $\mathbf{T}' = \mathbf{R}\mathbf{T}\mathbf{R}^T$.

Example. The gradient of a vector \mathbf{v}^a , denoted $\nabla\mathbf{v}$ is a second order cartesian tensor. Indeed, the components of $\nabla\mathbf{v}$ are:

$$(\nabla\mathbf{v})_{ij} = \frac{\partial v_i}{\partial x_j} \quad (49.2.31)$$

We may regard $\{\frac{\partial}{\partial x_j}\}$ as a basis, known as the **holonomic basis** or **coordinate basis**. Indeed:

$$\frac{\partial}{\partial x'_i} = \frac{\partial x_j}{\partial x'_i} \frac{\partial}{\partial x_j} = R_{ij} \frac{\partial}{\partial x_j} \implies R_{ij} = \frac{\partial x_j}{\partial x'_i} \quad (49.2.32)$$

from which it follows that ∇ is a first rank cartesian tensor^b.

Hence the components of $\nabla\mathbf{v}$ transform as:

$$(\nabla\mathbf{v})'_{ij} = \frac{\partial v'_i}{\partial x'_j} \quad (49.2.34)$$

$$= \frac{\partial v'_i}{\partial x_k} \frac{\partial x_k}{\partial x'_j} \quad (49.2.35)$$

$$= \frac{\partial}{\partial x_k} (R_{il}v_l) \frac{\partial x_k}{\partial x'_j} \quad (49.2.36)$$

$$= R_{il} \frac{\partial v_l}{\partial x_k} \frac{\partial x_k}{\partial x'_j} \quad (49.2.37)$$

$$= R_{il}R_{jk}(\nabla\mathbf{v})_{lk} \quad (49.2.38)$$

as would be expected from a second order tensor. ◀

^athis is a first order cartesian tensor

^bSimilarly:

$$R_{ji} = \frac{\partial x'_i}{\partial x_j} \quad (49.2.33)$$

Definition (Outer product)

The outer product of two vectors \mathbf{v} , \mathbf{u} is defined as:

$$(\mathbf{u} \otimes \mathbf{v})_{ij} = u_i v_j \quad (49.2.39)$$

and is a second order tensor.

It is easy to see that:

$$(\mathbf{u} \otimes \mathbf{v})'_{ij} = u'_i v'_j = R_{ik} R_{jl} u_k v_l = R_{ik} R_{jl} (\mathbf{u} \otimes \mathbf{v})_{kl} \quad (49.2.40)$$

as desired. Moreover, since $\mathbf{u} = u_i \mathbf{e}_i$ and $\mathbf{v} = v_i \mathbf{e}_i$ then:

$$\mathbf{u} \otimes \mathbf{v} = u_i v_j \mathbf{e}_i \otimes \mathbf{e}_j \quad (49.2.41)$$

where $\mathbf{e}_i \otimes \mathbf{e}_j$ is a sparse matrix with the only non-zero element $(\mathbf{e}_i \otimes \mathbf{e}_j)_{ij} = 1$.

Example. Consider the matrix:

$$\mathbf{T} = \begin{pmatrix} x_2^2 & -x_1 x_2 \\ -x_1 x_2 & x_1^2 \end{pmatrix} \quad (49.2.42)$$

Using $s \equiv \sin \theta$ and $c \equiv \cos \theta$ for shorthand we get that the components of \mathbf{T} transform as:

$$T'_{11} = (x'_2)^2 = (-x_1 s + x_2 c)^2 = x_1^2 s^2 + x_2^2 c^2 - 2x_1 x_2 c s \quad (49.2.43)$$

$$T'_{12} = -x'_1 x'_2 = -(x_1 c + x_2 s)(-x_1 s + x_2 c) = x_1^2 s c - x_2^2 s c + x_1 x_2 (s^2 - c^2) \quad (49.2.44)$$

$$T'_{21} = x_1^2 s c - x_2^2 s c + x_1 x_2 (s^2 - c^2) \quad (49.2.45)$$

$$T'_{22} = (x'_1)^2 = (x_1 c + x_2 s)^2 = x_1^2 c^2 + x_2^2 s^2 + 2x_1 x_2 s c \quad (49.2.46)$$

so that:

$$\mathbf{T}' = \begin{pmatrix} x_1^2 s^2 + x_2^2 c^2 - 2x_1 x_2 c s & x_1^2 s c - x_2^2 s c + x_1 x_2 (s^2 - c^2) \\ x_1^2 s c - x_2^2 s c + x_1 x_2 (s^2 - c^2) & x_1^2 c^2 + x_2^2 s^2 + 2x_1 x_2 s c \end{pmatrix} \quad (49.2.47)$$

If \mathbf{T} were a tensor then we would find

$$\mathbf{T}' = \mathbf{R} \mathbf{T} \mathbf{R}^T = \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} x_2^2 & -x_1 x_2 \\ -x_1 x_2 & x_1^2 \end{pmatrix} \begin{pmatrix} c & -s \\ s & c \end{pmatrix} \quad (49.2.48)$$

$$= \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} x_2^2 c - x_1 x_2 s & -x_2^2 s - x_1 x_2 c \\ -x_1 x_2 c + s^2 s & x_1 x_2 s + x_1^2 c \end{pmatrix} \quad (49.2.49)$$

$$= \begin{pmatrix} x_1^2 s^2 + x_2^2 c^2 - 2x_1 x_2 c s & x_1^2 s c - x_2^2 s c + x_1 x_2 (s^2 - c^2) \\ x_1^2 s c - x_2^2 s c + x_1 x_2 (s^2 - c^2) & x_1^2 c^2 + x_2^2 s^2 + 2x_1 x_2 s c \end{pmatrix} \quad (49.2.50)$$

so \mathbf{T} is indeed a second order cartesian tensor.

More simply, we could have noticed that $\mathbf{T} = (x_2, -x_1) \otimes (x_2, -x_1)$ and since $(x_2, -x_1)$ was proven to be a first order cartesian tensor, \mathbf{T} will be a second order cartesian tensor as desired. \blacktriangleleft

Definition (Cartesian Tensor) In general, a cartesian tensor is defined as a geometric object \mathbf{T} represented by the components $T_{ij...k}$ in the cartesian system \mathcal{C} and represented by the components $T'_{ij...k}$ in the cartesian system \mathcal{C}' , such that they transform under rigid cartesian rotations:

$$T'_{ij...k} = R_{ip}R_{jq}...R_{kr}T_{pq...r} \quad (49.2.51a)$$

$$T_{ij...k} = R_{pi}R_{qj}...R_{rk}T'_{pq...r} \quad (49.2.51b)$$

Theorem (Quotient law)

Suppose that \mathbf{B} and \mathbf{C} are tensors such that its components in any rotated basis satisfy:

$$A_{pq...k...m}B_{ij...k...n} = C_{pq...mij...n} \quad (49.2.52)$$

then \mathbf{A} must also be a tensor.

Proof. We consider the case for second order tensors (the general case follows exactly the same logic, just with more indices). We are given that:

$$A_{pk}B_{ik} = C_{pi}, \quad A'_{pk}B'_{ik} = C'_{pi} \quad (49.2.53)$$

and that $B_{jl} = R_{mj}R_{nl}B'_{mn}$, $C'_{pi} = R_{pq}R_{ij}C_{qj}$ then we get that:

$$A'_{pk}B'_{ik} = C'_{pi} \quad (49.2.54)$$

$$= R_{pq}R_{ij}C_{qj} \quad (49.2.55)$$

$$= R_{pq}R_{ij}A_{ql}B_{jl} \quad (49.2.56)$$

$$= R_{pq}R_{ij}A_{ql}R_{mj}R_{nl}B'_{mn} \quad (49.2.57)$$

$$= R_{pq}R_{nl}A_{ql}B'_{in} \quad (49.2.58)$$

so that:

$$(A'_{pk} - R_{pq}R_{nl}A_{ql})B'_{ik} = 0 \quad (49.2.59)$$

Since this must hold for any B'_{ik} we must have that:

$$A'_{pk} = R_{pq}R_{nl}A_{ql} \quad (49.2.60)$$

as desired. ■

The quotient law is a much faster way to prove that a quantity is a tensor, since it suffices to contract this quantity with a known tensor and ensure that the resulting quantity is also a tensor.

Example. Let's prove that:

$$\mathbf{T} = \begin{pmatrix} x_2^2 & -x_1 x_2 \\ -x_1 x_2 & x_1^2 \end{pmatrix} \quad (49.2.61)$$

is a tensor. We have already done this in two ways, but a third way is by using the quotient law:

$$T_{11}x_1^2 = x_1^2 x_2^2 \quad (49.2.62)$$

$$T_{12}x_1 x_2 = -x_1^2 x_2^2 \quad (49.2.63)$$

$$T_{21}x_2 x_1 = -x_1^2 x_2^2 \quad (49.2.64)$$

$$T_{22}x_2^2 = x_1^2 x_2^2 \quad (49.2.65)$$

so that $T_{ij}x_i x_j = 0$ which is a tensor. Since $x_i x_j$ is an outer product and thus a tensor, it follows that \mathbf{T} is also a tensor. \blacktriangleleft

49.3 The δ_{ij} and ϵ_{ijk} tensors

49.4 Physical examples of cartesian tensors

Consider the angular momentum $\mathbf{L} = \mathbf{r} \times \mathbf{p} = m\mathbf{r} \times \dot{\mathbf{r}}$, this is a first order cartesian tensor.

Indeed if we write the components of \mathbf{L} as:

$$L_i = m\epsilon_{ijk}r_j \dot{r}_k \quad (49.4.1)$$

they they will transform under rotations as:

$$L'_i = m\epsilon_{ijk}r'_j \dot{r}'_k \quad (49.4.2)$$

and since $r'_j = R_{jm}r_m$ and $\dot{r}'_k = \frac{d}{dt}(R_{kn}r_n) = R_{kn}\dot{r}_n$ we will find that.

$$L'_i = m\epsilon_{ijk}R_{jm}R_{kn}r_m \dot{r}_n \quad (49.4.3)$$

Consider the i, j component of the **inertia tensor**:

$$I_{ij} = \int (\delta_{ij}r^2 - x_i x_j) dm \quad (49.4.4)$$

In dirac notation, it is easy to see that $I_{ij} = \langle i|I|j\rangle$, $\delta_{ij} = \langle i|j\rangle$, $r^2 = \langle r|r\rangle$, $x_i = \langle r|i\rangle = \langle i|r\rangle$. Consequently:

$$\langle i|I|j\rangle = \langle i| \int \langle r|r\rangle \mathbb{I} - |r\rangle \langle r| dm |j\rangle \quad (49.4.5)$$

from which it follows that:

$$I = \int r^2 \mathbb{I} - dm \quad (49.4.6)$$

This is a second order cartesian tensor.

49.5 Non-cartesian tensors

Definition (*Contravariant and covariant bases*)

We saw that for general curvilinear coordinates (u_1, u_2, u_3) , so that a position vector may be expressed as $\mathbf{r}(u_1, u_2, u_3)$ then we have two important sets of basis vectors:

$$\mathbf{e}_i = \frac{\partial \mathbf{r}}{\partial u^i}, \quad \mathbf{e}^i = \nabla u_i \quad (49.5.1)$$

We have slightly modified our summation convention to include superscripts. From now on, we will assume that any lower case index that appears exactly once as a superscript and once as a subscript must be summed over.

Proposition (*Reciprocity relation*)

The contravariant and covariant bases are orthonormal to each other:

$$\mathbf{e}_i \cdot \mathbf{e}^j = \delta_i^j \quad (49.5.2)$$

Proof. We find that:

$$\mathbf{e}_i \cdot \mathbf{e}^j = \frac{\partial \mathbf{r}}{\partial u^i} \cdot \nabla u_j \quad (49.5.3)$$

$$= \frac{\partial u_1}{\partial u^i} \frac{\partial u_j}{\partial u_1} + \frac{\partial u_2}{\partial u^i} \frac{\partial u_j}{\partial u_2} + \frac{\partial u_3}{\partial u^i} \frac{\partial u_j}{\partial u_3} \quad (49.5.4)$$

$$= \frac{\partial u_j}{\partial u^i} = \delta_i^j \quad (49.5.5)$$

as desired. ■

Consequently, given some vector \mathbf{a} it may be expanded in both the contravariant and covariant bases. We find that if

$$\mathbf{a} = a^i \mathbf{e}_i = a_i \mathbf{e}^i \quad (49.5.6)$$

where a^i are the contravariant components while a_i are the covariant components, then:

$$\mathbf{a} \cdot \mathbf{e}^j = a^i \delta_i^j = a^j \quad (49.5.7)$$

$$\mathbf{a} \cdot \mathbf{e}_j = a_i \delta_j^i = a_j \quad (49.5.8)$$

as expected.

Let's now consider an infinitesimal vector displacement $d\mathbf{r} = du^i \mathbf{e}_i$. Then, the infinitesimal arc length is:

$$(ds)^2 = du^i du^j \mathbf{e}_i \cdot \mathbf{e}_j = g_{ij} du^i du^j \quad (49.5.9)$$

where $g_{ij} \equiv \mathbf{e}_i \cdot \mathbf{e}_j$ is defined as the metric tensor.

Definition (Metric tensor)

For a given set of curvilinear coordinates (u_i) with contravariant and covariant bases $\{\mathbf{e}_i = \frac{\partial \mathbf{r}}{\partial u_i}\}$ and $\{\mathbf{e}^i = \nabla u_i\}$ respectively, the metric tensor is defined to be:

$$g_{ij} \equiv \mathbf{e}_i \cdot \mathbf{e}_j \quad (49.5.10)$$

Furthermore, the volume element may be expressed as:

$$dV = \sqrt{g} du^i \quad (49.5.11)$$

where $g = \det[g_{ij}]$ is the determinant of the metric tensor.

Example. Let's evaluate the metric tensor in spherical polar coordinates. We have that the position vector of some general point $(u_1, u_2, u_3) = (r, \theta, \phi)$ is given by:

$$\mathbf{r} = r \sin \theta \cos \phi \mathbf{e}_x + r \sin \theta \sin \phi \mathbf{e}_y + r \cos \theta \mathbf{e}_z \quad (49.5.12)$$

The covariant basis is easily found to be:

$$\mathbf{e}_1 = \frac{\partial \mathbf{r}}{\partial r} = \sin \theta \cos \phi \mathbf{e}_x + \sin \theta \sin \phi \mathbf{e}_y + \cos \theta \mathbf{e}_z \quad (49.5.13)$$

$$\mathbf{e}_2 = \frac{\partial \mathbf{r}}{\partial \theta} = r \cos \theta \cos \phi \mathbf{e}_x + r \cos \theta \sin \phi \mathbf{e}_y - r \sin \theta \mathbf{e}_z \quad (49.5.14)$$

$$\mathbf{e}_3 = \frac{\partial \mathbf{r}}{\partial \phi} = -r \sin \theta \sin \phi \mathbf{e}_x + r \sin \theta \cos \phi \mathbf{e}_y \quad (49.5.15)$$

Consequently:

$$[g_{ij}] = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2 \sin^2 \theta \end{pmatrix} \quad (49.5.16)$$

As expected from an orthogonal coordinate system, the metric is diagonal. Moreover, we find that the infinitesimal arc length may be expressed as

$$(ds)^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \quad (49.5.17)$$

while the volume element is found to be:

$$dV = r^2 \sin \theta dr d\theta d\phi \quad (49.5.18)$$



Just like vectors, tensors may also be expressed in both bases:

$$\mathbf{T} = T^{ij} \mathbf{e}_i \otimes \mathbf{e}_j = T_{ij} \mathbf{e}^i \otimes \mathbf{e}^j = T_i^j \mathbf{e}^i \otimes \mathbf{e}_j \quad (49.5.19)$$

Here T^{ij} , T_{ij} , T_i^j are known as the contravariant, covariant and mixed tensor components respectively.

We can use the metric tensor to express the scalar product of two vectors. Indeed, using the covariant and contravariant expansions:

$$\mathbf{a} \cdot \mathbf{b} = a^i \mathbf{e}_i b^j \mathbf{e}_j = g_{ij} a^i b^j = a_i \mathbf{e}^i b_j \mathbf{e}^j = g^{ij} a_i b_j \quad (49.5.20)$$

Finally, we also find that:

$$\mathbf{a} \cdot \mathbf{b} = a^i \mathbf{e}_i b_j \mathbf{e}^j = a^i b_j \delta_i^j = a^i b_i = a_i b^i \quad (49.5.21)$$

Consequently:

$$g_{ij} a^i b^j = a^i b_i, \quad g_{ij} a_i b_j = a_i b^i \quad (49.5.22)$$

This holds for any arbitrary \mathbf{a} , so we get the following very useful result:

Theorem (Raising and lowering indices)

For a given vector \mathbf{b} expressed in a set of curvilinear coordinates with metric tensor g_{ij} , its covariant and contravariant components are related by:

$$g_{ij} b^j = b_i, \quad g^{ij} b_j = b^i \quad (49.5.23)$$

The special case when $\mathbf{b} = \mathbf{e}_i$ then:

$$\mathbf{e}^i = g^{ij} \mathbf{e}_j, \quad \mathbf{e}_i = g_{ij} \mathbf{e}^j \quad (49.5.24)$$

Proposition (Contravariant and covariant components are inverses)

The contravariant g^{ij} and covariant g_{ij} components of a metric tensor obey:

$$g^{ij} g_{jk} = \delta_k^i \quad (49.5.25)$$

Proof. Consider:

$$a^i = \delta_k^i a^k = g^{ij} a_j = g^{ij} g_{jk} a^k \implies \delta_k^i = g^{ij} g_{jk} \quad (49.5.26)$$

as desired. ■

49.6 Covariance and contravariance

Let's now see what happens how covariant and contravariant components transform when we perform a change of curvilinear coordinates, since this is how we initially defined a tensor.

Suppose we have a coordinate system $\{u^i\}$ which we transform to another system $\{u'^i\}$. The new sets of basis vectors are:

$$\mathbf{e}'_i = \frac{\partial \mathbf{r}}{\partial u'^i}, \quad \mathbf{e}^{i'} = \nabla u^{i'} \quad (49.6.1)$$

The new covariant basis can be related to the old one by:

$$\mathbf{e}'_i = \frac{\partial \mathbf{r}}{\partial u'^i} = \frac{\partial u^j}{\partial u'^i} \frac{\partial \mathbf{r}}{\partial u^j} = \frac{\partial u^j}{\partial u'^i} \mathbf{e}_j \quad (49.6.2)$$

or alternatively:

$$\mathbf{e}_j = \frac{\partial u'^i}{\partial u^j} \mathbf{e}'_i \quad (49.6.3)$$

Therefore, expanding an arbitrary vector in both covariant bases:

$$\mathbf{a} = a^i \mathbf{e}_i = a^i \frac{\partial u'^j}{\partial u^i} \mathbf{e}'_j = a'^j \mathbf{e}'_j \quad (49.6.4)$$

implying that:

$$a'^j = \frac{\partial u'^j}{\partial u^i} a^i \quad (49.6.5)$$

Similarly, the new contravariant basis can be related to the old one by:

$$\mathbf{e}^{i'} = \nabla u'^i = \frac{\partial u'^i}{\partial u^j} \nabla u^j = \frac{\partial u'^i}{\partial u^j} \mathbf{e}^j \quad (49.6.6)$$

or alternatively:

$$\mathbf{e}^i = \frac{\partial u^i}{\partial u'^j} \mathbf{e}'^j \quad (49.6.7)$$

Therefore, expanding an arbitrary vector in both contravariant bases we find that:

$$\mathbf{a} = a_i \mathbf{e}^i = a_i \frac{\partial u^i}{\partial u'^j} \mathbf{e}'^j = a'_j \mathbf{e}'^j \quad (49.6.8)$$

implying that covariant components transform as:

$$a'_j = \frac{\partial u^i}{\partial u'^j} a_i \quad (49.6.9)$$

Theorem (Transformation of co(ntra)variant components) For a given vector \mathbf{a} , its covariant and contravariant components transform under a change of basis as:

$$a'^j = \frac{\partial u'^j}{\partial u^i} a^i \quad (49.6.10)$$

$$a'_j = \frac{\partial u^i}{\partial u'^j} a_i \quad (49.6.11)$$

In a completely analogous way, we can show how the co(ntra)variant and mixed components of a second rank tensor transform:

$$T'^{ij} = \frac{\partial u'^i}{\partial u^k} \frac{\partial u'^j}{\partial u^l} T^{kl} \quad (49.6.12)$$

$$T'^i_j = \frac{\partial u'^i}{\partial u^k} \frac{\partial u^l}{\partial u'^j} T^k_l \quad (49.6.13)$$

$$T'_{ij} = \frac{\partial u^k}{\partial u'^i} \frac{\partial u^l}{\partial u'^j} T_{kl} \quad (49.6.14)$$

Proof. We have already proven the result for vectors. For second rank tensors, we find:

$$T'^{ij} \mathbf{e}'_i \otimes \mathbf{e}'_j = T^{ij} \mathbf{e}_i \otimes \mathbf{e}_j \quad (49.6.15)$$

$$T'^{ij} \mathbf{e}^{i'} \otimes \mathbf{e}'_j = T^j_i \mathbf{e}^i \otimes \mathbf{e}_j \quad (49.6.16)$$

$$T'_{ij} \mathbf{e}^{i'} \otimes \mathbf{e}^{j'} = T_{ij} \mathbf{e}^i \otimes \mathbf{e}^j \quad (49.6.17)$$

For example, we would find that:

$$T'^{ij} \mathbf{e}'_i \otimes \mathbf{e}'_j = T^{kl} \mathbf{e}_k \otimes \mathbf{e}_l = T^{kl} \frac{\partial u'^i}{\partial u^k} \frac{\partial u'^j}{\partial u^l} \mathbf{e}'_i \otimes \mathbf{e}'_j \quad (49.6.18)$$

implying

$$T'^{ij} = \frac{\partial u'^i}{\partial u^k} \frac{\partial u'^j}{\partial u^l} T^{kl} \quad (49.6.19)$$

as desired. ■

Visually, we can explain covariant and contravariant components as follows: “contravariant components transform as position vector components, while covariant components transform as gradient vector components”. This makes physically sense, as the contravariant basis is parallel everywhere to its coordinate curves, while the covariant basis is orthonormal everywhere to its coordinate surfaces. This is in alignment with the intuition that

coordinate curves transform as position vectors, while coordinate surfaces transform as gradient vectors.

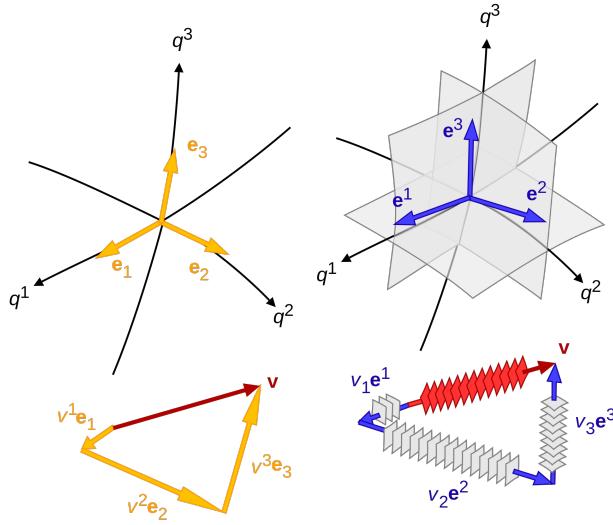


Figure 49.1. By Maschen - Own work, CC0, <https://commons.wikimedia.org/w/index.php?curid=21043814>

Clearly, if we increase the lengths of the contravariant basis vectors, the contravariant components decrease (less arrows along each basis vector are needed), and vice versa. The contravariant components “contra-vary” with the change of basis.

On the other hand, if I increase the lengths of the covariant basis vectors, the covariant components increase (“more” planes can be packed along each basis vector), and vice versa. The covariant components “co-vary” with the change of basis.

Theorem (Metric tensor is a second rank tensor)

The metric tensor $g_{\mu\nu}$ is a second-rank tensor.

Proof. We have that :

$$g_{\mu\nu}dx^\mu dx^\nu = g'_{\alpha\beta}dx'^\alpha dx'^\beta \quad (49.6.20)$$

$$= g'_{\alpha\beta} \frac{\partial x'^\alpha}{\partial x^\mu} \frac{\partial x'^\beta}{\partial x^\nu} dx^\mu dx^\nu \quad (49.6.21)$$

$$\Rightarrow \left(g_{\mu\nu} - g'_{\alpha\beta} \frac{\partial x'^\alpha}{\partial x^\mu} \frac{\partial x'^\beta}{\partial x^\nu} \right) dx^\mu dx^\nu = 0 \quad (49.6.22)$$

Seeing as this must hold for all dx^μ, dx^ν , it follows that:

$$g_{\mu\nu} = \frac{\partial x'^\alpha}{\partial x^\mu} \frac{\partial x'^\beta}{\partial x^\nu} g'_{\alpha\beta} \quad (49.6.23)$$

proving the required transformation rule. ■

49.7 Application to special relativity: four-vectors

In special relativity we deal with four vectors whose components transform like:

$$v'^\mu = \frac{\partial x'^\mu}{\partial x^\nu} v^\nu \quad (49.7.1)$$

where $\frac{\partial x'^\mu}{\partial x^\nu}$ are components of the Lorentz transformation. An important four vector is the displacement vector:

$$dx^\mu = (cdt, dx, dy, dz) = (dx^0, dx^1, dx^2, dx^3) = (cdt, d\mathbf{r}) \quad (49.7.2)$$

We also know that the following quantity, known as the proper time, is an invariant under Lorentz transformations:

$$(cd\tau)^2 = c^2 dt^2 - d\mathbf{r} \cdot d\mathbf{r} \quad (49.7.3)$$

This suggests introducing the following metric tensor, known as the Minkowski metric:

$$[\eta_{\mu\nu}] = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \implies (cd\tau)^2 = g_{\mu\nu} dx^\mu dx^\nu \quad (49.7.4)$$

This allows us to find the covariant components:

$$dx_\mu = (cdt, -d\mathbf{r}) \quad (49.7.5)$$

The importance of the proper time arises when we try to define a velocity four vector. Suppose we naively define it to be:

$$u^\mu = \frac{\partial x^\mu}{\partial t} = \left(c, \frac{d\mathbf{r}}{dt} \right) \quad (49.7.6)$$

This result is quite worrisome, since the first component c of this supposed four vector is constant, and therefore does not transform at all under a change of basis.

Suppose we instead define

$$u^\mu = \frac{\partial x^\mu}{\partial \tau} = \frac{\partial x^\mu}{\partial t} \frac{\partial t}{\partial \tau} = \left(c, \frac{d\mathbf{r}}{dt} \right) \frac{dt}{d\tau} \quad (49.7.7)$$

Then, since

$$c^2 d\tau^2 = c^2 dt^2 - d\mathbf{r} \cdot d\mathbf{r} = dt^2 \left(c^2 - \frac{d\mathbf{r}}{dt} \cdot \frac{d\mathbf{r}}{dt} \right) \quad (49.7.8)$$

$$\implies \frac{dt}{d\tau} = \frac{1}{\sqrt{1 - v^2/c^2}} \equiv \gamma \quad (49.7.9)$$

where we defined $\mathbf{v} \equiv \frac{d\mathbf{r}}{dt}$, we find that the velocity four vector reads:

$$u^\mu = \gamma(c, \mathbf{v}) \quad (49.7.10)$$

and thus:

$$u_\mu = \gamma(c, -\mathbf{v}) \quad (49.7.11)$$

Tensor calculus

50.1 Christoffel symbols

Let's consider how we can take derivatives of basis vectors. For example, consider $\frac{\partial \mathbf{e}_i}{\partial u^j}$. Since this is itself a vector, it can be expanded in the covariant basis as::

$$\boxed{\frac{\partial \mathbf{e}_i}{\partial u^j} = \Gamma_{ij}^k \mathbf{e}_k} \quad (50.1.1)$$

It is easy to see that:

$$\Gamma_{ij}^k = \mathbf{e}^k \cdot \frac{\partial \mathbf{e}_i}{\partial u^j} \quad (50.1.2)$$

Also, we can differentiate the relation $\mathbf{e}^i \cdot \mathbf{e}_j$ to find:

$$\frac{\partial}{\partial u^k} (\mathbf{e}^i \cdot \mathbf{e}_j) = \frac{\partial \mathbf{e}^i}{\partial u^k} \cdot \mathbf{e}_j + \mathbf{e}^i \cdot \frac{\partial \mathbf{e}_j}{\partial u^k} = 0 \quad (50.1.3)$$

$$\Rightarrow \frac{\partial \mathbf{e}^i}{\partial u^k} \cdot \mathbf{e}_j = -\Gamma_{jk}^i \quad (50.1.4)$$

$$\boxed{\Rightarrow \frac{\partial \mathbf{e}^i}{\partial u^j} = -\Gamma_{kj}^i \mathbf{e}^k} \quad (50.1.5)$$

The Γ_{ij}^k are known as Christoffel symbols. Although it looks like a third rank tensor, it actually does not follow the required transformation laws. Indeed:

$$\Gamma'_{ij}^k = \mathbf{e}'^k \cdot \frac{\partial \mathbf{e}'_i}{\partial u'^j} \quad (50.1.6)$$

$$= \frac{\partial u'^k}{\partial u^n} \mathbf{e}^n \cdot \frac{\partial}{\partial u'^k} \left(\frac{\partial u^l}{\partial u'^i} \mathbf{e}_l \right) \quad (50.1.7)$$

$$= \frac{\partial u'^k}{\partial u^n} \mathbf{e}^n \cdot \left(\frac{\partial^2 u^l}{\partial u'^j \partial u'^i} \mathbf{e}_l + \frac{\partial u^l}{\partial u'^i} \frac{\partial \mathbf{e}_l}{\partial u^m} \frac{\partial u^m}{\partial u'^j} \right) \quad (50.1.8)$$

$$= \frac{\partial u'^k}{\partial u^n} \frac{\partial^2 u^l}{\partial u'^j \partial u'^i} \delta_l^n + \frac{\partial u'^k}{\partial u^n} \frac{\partial u^l}{\partial u'^i} \frac{\partial u^m}{\partial u'^j} \mathbf{e}_n \cdot \frac{\partial \mathbf{e}_l}{\partial u^m} \quad (50.1.9)$$

$$= \frac{\partial u'^k}{\partial u^l} \frac{\partial^2 u^l}{\partial u'^j \partial u'^i} + \Gamma_{lm}^n \frac{\partial u'^k}{\partial u^n} \frac{\partial u^l}{\partial u'^i} \frac{\partial u^m}{\partial u'^j} \quad (50.1.10)$$

We can find another expression for the Christoffel symbols that allow for faster computation. Consider the derivative of the metric tensor:

$$\frac{dg_{ij}}{du^k} = \frac{\partial \mathbf{e}_i}{\partial u^k} \cdot \mathbf{e}_j + \mathbf{e}_i \cdot \frac{\partial \mathbf{e}_j}{\partial u^k} \quad (50.1.11)$$

$$= \Gamma_{ik}^l \mathbf{e}_l \cdot \mathbf{e}_j + \mathbf{e}_i \cdot \mathbf{e}_l \Gamma_{jk}^l \quad (50.1.12)$$

$$\implies \frac{dg_{ij}}{du^k} = \Gamma_{ik}^l g_{lj} + \Gamma_{jk}^l g_{il} \quad (50.1.13)$$

Now note that Γ_{ik}^l is symmetric, since:

$$\frac{\partial \mathbf{e}_i}{\partial u^j} = \frac{\partial^2 \mathbf{r}}{\partial u^i \partial u^j} = \frac{\partial \mathbf{e}_j}{\partial u^i} \quad (50.1.14)$$

Consequently, we can simply permute the indices in (50.1.13) and find the following:

$$\frac{dg_{ik}}{du^j} = \Gamma_{ij}^l g_{lk} + \Gamma_{jk}^l g_{il} \quad (50.1.15)$$

$$\frac{dg_{kj}}{du^i} = \Gamma_{ik}^l g_{lj} + \Gamma_{ji}^l g_{kl} \quad (50.1.16)$$

implying that:

$$\frac{dg_{ij}}{du^k} + \frac{dg_{ik}}{du^j} - \frac{dg_{kj}}{du^i} = 2\Gamma_{jk}^l g_{il} \quad (50.1.17)$$

$$\iff \left(\frac{dg_{ij}}{du^k} + \frac{dg_{ik}}{du^j} - \frac{dg_{kj}}{du^i} \right) g^{im} = 2\Gamma_{jk}^l g_{il} g^{im} \quad (50.1.18)$$

$$\boxed{\iff \Gamma_{jk}^m = \frac{1}{2} g^{im} \left(\frac{dg_{ij}}{du^k} + \frac{dg_{ik}}{du^j} - \frac{dg_{kj}}{du^i} \right)} \quad (50.1.19)$$

50.2 Differentiating tensors

Let's consider a contravariant vector:

$$x'^i = \frac{\partial x'^i}{\partial x^j} x^j \quad (50.2.1)$$

Its time derivative is transforms following:

$$\frac{\partial x'^i}{\partial u'^j} = \frac{\partial u^k}{\partial u'^j} \frac{\partial x'^i}{\partial u^k} \quad (50.2.2)$$

$$= \frac{\partial u^k}{\partial u'^j} \frac{\partial}{\partial u^k} \left(\frac{\partial x'^i}{\partial u^l} x^l \right) \quad (50.2.3)$$

$$= \frac{\partial u^k}{\partial u'^j} \left(\frac{\partial^2 u'^i}{\partial u^l \partial u^k} x^l + \frac{\partial u'^i}{\partial u^l} \frac{\partial x^l}{\partial u^k} \right) \quad (50.2.4)$$

$$= \frac{\partial u^k}{\partial u'^j} \frac{\partial^2 u'^i}{\partial u^j \partial u^k} x^j + \frac{\partial u^k}{\partial u'^j} \frac{\partial u'^i}{\partial u^l} \frac{\partial x^l}{\partial u^k} \quad (50.2.5)$$

The $\frac{\partial u^k}{\partial u^j} \frac{\partial^2 u^i}{\partial u^j \partial u^k} x^j$ term ruins the transformation rule, and if it is non-zero then $\frac{\partial x^i}{\partial x^j}$ cannot be considered a tensor. This is quite problematic, as from our experience things such as the velocity vector should be a tensor.

Luckily, we can use the Christoffel symbols to deal with this issue. Consider:

$$\frac{\partial \mathbf{x}}{\partial u^j} = \frac{\partial x^i}{\partial u^j} \mathbf{e}_i + x^i \Gamma_{ij}^k \mathbf{e}_k \quad (50.2.6)$$

$$= \left(\frac{\partial x^i}{\partial u^j} + x^k \Gamma_{kj}^i \right) \mathbf{e}_i \quad (50.2.7)$$

leading us to defining the covariant derivative as follows:

Definition (Covariant derivative)

Given a contravariant representation of a vector $\mathbf{v} = v^i \mathbf{e}_i$, its **covariant derivative** is defined as:

$$v_{:j}^i = \frac{\partial x^i}{\partial u^j} + x^k \Gamma_{kj}^i \quad (50.2.8)$$

and similarly for the covariant representation:

$$v_{i:j} = \frac{\partial x_i}{\partial u^j} - x_k \Gamma_{ij}^k \quad (50.2.9)$$

We see that the covariant derivative is different from the normal partial derivative in that the basis vectors that don't change over space for cartesian coordinates are variable in general coordinates. Due to the product derivative rule, this creates an extra term in the derivative which we identify using a Christoffel symbol.

Example. Let's work out the covariant derivative of the contravariant components of a second order tensor \mathbb{T} .

The contravariant components T^{ij} satisfy:

$$\mathbb{T} = T^{ij} (\mathbf{e}_i \otimes \mathbf{e}_j) \quad (50.2.10)$$

Consequently:

$$\frac{\partial \mathbb{T}}{\partial u^k} = \frac{\partial T^{ij}}{\partial u^k} (\mathbf{e}_i \otimes \mathbf{e}_j) + T^{ij} \frac{\partial}{\partial u^k} (\mathbf{e}_i \otimes \mathbf{e}_j) \quad (50.2.11)$$

We can simplify the second term on the RHS:

$$\frac{\partial}{\partial u^k} (\mathbf{e}_i \otimes \mathbf{e}_j) = \frac{\partial \mathbf{e}_i}{\partial u^k} \otimes \mathbf{e}_j + \mathbf{e}_i \otimes \frac{\partial \mathbf{e}_j}{\partial u^k} \quad (50.2.12)$$

$$= \Gamma_{ik}^l \mathbf{e}_l \otimes \mathbf{e}_j + \mathbf{e}_i \otimes \mathbf{e}_l \Gamma_{jk}^l \quad (50.2.13)$$

giving

$$\frac{\partial \mathbb{T}}{\partial u^k} = \frac{\partial T^{ij}}{\partial u^k} (\mathbf{e}_i \otimes \mathbf{e}_j) + T^{ij} (\Gamma_{ik}^l \mathbf{e}_l \otimes \mathbf{e}_j + \mathbf{e}_i \otimes \mathbf{e}_l \Gamma_{jk}^l) \quad (50.2.14)$$

$$= \left(\frac{\partial T^{ij}}{\partial u^k} + T^{lj} \Gamma_{lk}^i + T^{il} \Gamma_{lk}^j \right) (\mathbf{e}_i \otimes \mathbf{e}_j) \quad (50.2.15)$$

$$= T_{:k}^{ij} (\mathbf{e}_i \otimes \mathbf{e}_j) \quad (50.2.16)$$

where we defined a covariant derivative:

$$T_{:k}^{ij} \equiv \frac{\partial T^{ij}}{\partial u^k} + T^{lj} \Gamma_{lk}^i + T^{il} \Gamma_{lk}^j \quad (50.2.17)$$



50.3 Application to geometry: curvilinear coordinates

Let's apply our knowledge of tensor calculus to study general curvilinear coordinates.

Gradient

The gradient of a scalar field ϕ is just:

$$\nabla \phi \equiv \phi_{;i} \mathbf{e}^i = \frac{\partial \phi}{\partial u^i} \mathbf{e}^i \quad (50.3.1)$$

Divergence

The divergence of a vector field \mathbf{v} is:

$$\nabla \cdot \mathbf{x} \equiv v_{:i}^i = \frac{\partial v^i}{\partial u^i} + x^k \Gamma_{ki}^i \quad (50.3.2)$$

Note that the Christoffel symbol simplifies significantly:

$$\Gamma_{ki}^i = \frac{1}{2} g^{il} \left(\frac{\partial g_{il}}{\partial u^k} + \frac{\partial g_{kl}}{\partial u^i} - \frac{\partial g_{ki}}{\partial u^l} \right) \quad (50.3.3)$$

$$= \frac{1}{2} g^{il} \left(\frac{\partial g_{il}}{\partial u^k} + \frac{\partial g_{kl}}{\partial u^i} - \frac{\partial g_{kl}}{\partial u^i} \right) \quad (50.3.4)$$

$$= \frac{1}{2} g^{il} \frac{\partial g_{il}}{\partial u^k} \quad (50.3.5)$$

giving:

$$\nabla \cdot \mathbf{x} = \frac{\partial v^i}{\partial u^i} + \frac{1}{2} x^k g^{il} \frac{\partial g_{il}}{\partial u^k} \quad (50.3.6)$$

Proposition (Important determinant identity)

Let $\mathbf{A} = (a_{ij})$, $\mathbf{B} = (b^{ij})$ and $\mathbf{A} = \mathbf{B}^{-1}$. Then:

$$\frac{\partial |\mathbf{A}|}{\partial u^k} = |\mathbf{A}| b^{ji} \frac{\partial a_{ij}}{\partial u^k} \quad (50.3.7)$$

Proof. We have that if the cofactor of the element a_{ij} is c^{ij} then:

$$b^{ij} = \frac{1}{|\mathbf{A}|} c^{ji} \quad (50.3.8)$$

Consequently, since $|\mathbf{A}| = a_{ij} c^{ij}$ with fixed i :

$$\frac{\partial |\mathbf{A}|}{\partial a_{ij}} = \frac{\partial |\mathbf{A}|}{\partial a_{ij}} \frac{\partial a_{ij}}{\partial u^k} = |\mathbf{A}| b^{ij} \frac{\partial a_{ij}}{\partial u^k} \quad (50.3.9)$$

implying:

$$\frac{\partial |\mathbf{A}|}{\partial u^k} = \frac{\partial |\mathbf{A}|}{\partial a_{ij}} \frac{\partial a_{ij}}{\partial u^k} = |\mathbf{A}| b^{ij} \frac{\partial a_{ij}}{\partial u^k} \quad (50.3.10)$$

as desired. ■

We can apply this proposition to the determinant of the metric tensor:

$$\frac{\partial |g|}{\partial u^k} = |g| g^{ij} \frac{\partial g_{ij}}{\partial u^k} \implies g^{ij} \frac{\partial g_{ij}}{\partial u^k} = \frac{1}{|g|} \frac{\partial |g|}{\partial u^k} = \frac{1}{\sqrt{|g|}} \frac{\partial \sqrt{|g|}}{\partial u^k} \quad (50.3.11)$$

and hence when substituted into (50.3.6) we get:

$$\nabla \cdot \mathbf{x} = \frac{\partial v^i}{\partial u^i} + \frac{1}{\sqrt{|g|}} \frac{\partial \sqrt{|g|}}{\partial u^i} x^i \quad (50.3.12)$$

or alternatively:

$$\boxed{\nabla \cdot \mathbf{v} = \frac{1}{\sqrt{|g|}} \frac{\partial (\sqrt{|g|} x^i)}{\partial u^i}} \quad (50.3.13)$$

Laplacian
Curl

50.4 Geodesics

Part VI

Differential Geometry

Differentiable Manifolds

51

Differential forms

52

Integrating on manifolds

53

Curvature

54

Lie derivatives

55

Part VII

Complex analysis

Part VIII

Calculus of Variations

Part IX

Fourier Analysis

Part X

Functional Analysis and Operator theory

Acknowledgments

This is the most common positions for acknowledgments. A macro is available to maintain the same layout and spelling of the heading.

Note added. This is also a good position for notes added after the paper has been written.

Bibliography

- [1] Author, *Title*, *J. Abbrev.* **vol** (year) pg.
- [2] Author, *Title*, arxiv:1234.5678.
- [3] Author, *Title*, Publisher (year).